Weakly Supervised Temporal Modeling of Latent Dynamics in Dyadic Conversations

Tahiya Chowdhury
Department of Computer Science & Davis Institute for AI
Colby College
Waterville, Maine, USA
tahiya.chowdhury@colby.edu

Abstract

Conversations in collaborative settings involve evolving social and cognitive dynamics such as coordination, cognitive load, and participation asymmetry, which are not directly observable but manifest through interactional behavior over time. Prior work has largely relied on static summaries or post-hoc labels, limiting our ability to capture how these dynamics unfold during interaction. We model dyadic conversation as a partially observable temporal process and propose a weakly supervised framework for tracking latent conversational state from interaction and acoustic behavioral signals. Using a dataset of remote dyadic conversations (53 dyads) over 9 collaborative tasks with task-level annotations, we segment interactions into fixed temporal windows of 30 seconds and extract 34 features capturing interactional features (turn-taking, floor control) and speaker-relative acoustic measures. We compare static models (Ridge, Random Forest), sequential neural models (GRU with pooling and attention), and two strategies for temporal trajectory modeling of latent states. Static interaction features remain the strongest predictor for both temporal demand (correlation $= 0.254$) and mental demand (correlation $= 0.217$); but we do not claim temporal modeling improves prediction accuracy over static baselines. More importantly, temporal modeling reveals interpretable latent trajectory structures – escalation, convergence, and participation imbalance, which are not observable in aggregated summary features alone and can vary systematically by task type and cognitive demand level. These findings provide a step toward dynamic, interpretable representations of conversational state in human conversations.

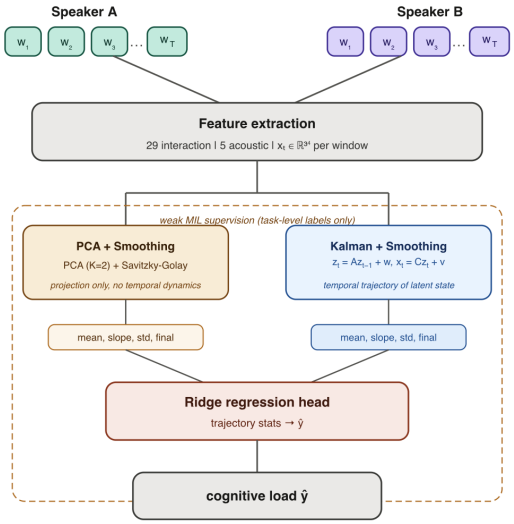


Figure 1: Weakly supervised temporal modeling pipeline. Speaker-separated fixed lengths windows feed joint feature extraction, which then go through two latent trajectory estimation: PCA and Kalman smoothing. The trajectory statistics predict the task-level cognitive load for the dyad under weak supervision.

1 Introduction

Remote collaborative work increasingly relies on voice-mediated dyadic interaction, where participants must coordinate cognitively, manage conversational floor control, and adapt to each other's communicative behavior in real time. Understanding the latent social-cognitive state of a conversation—the unobservable conditions of cognitive load, coordination quality, and participation asymmetry that shape how people interact—is a fundamental challenge in dialogue modeling.

Prior work on dialogue state tracking (DST) has predominantly focused on slot-value semantics in task-oriented systems (Young et al., 2013; Williams and Young, 2007), treating state as a discrete, annotated outcome updated at each turn. Research on social signals and cognitive load in human–human conversation has characterized behavioral correlates of load from speech and video (Vinciarelli et al., 2009; Vukovic et al., 2021; Yang et al., 2023), but has largely adopted post-hoc, static approaches:

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 440–452
August 2–5, 2026.

collect task-level self-report labels and predict them from aggregate conversation features. Both traditions treat conversational state as a fixed endpoint rather than an evolving process that changes over time.

We argue that collaborative dyadic conversation should be modeled as a *partially observable dynamic system*. The social-cognitive state of a dyad at any moment is not directly observable, and it changes *within* a task as the interaction unfolds. This temporal evolution carries information that static summaries discard.

In this work, we formalize this idea and follow prior work (Williams and Young, 2007; Young et al., 2013, 2010) modeling dyadic conversations as a Partially Observable Markov Decision Process to propose a weakly supervised framework for tracking latent conversational state. Because labels are generally collected at the task level through self-reporting with no within-task annotations, our setting constitutes a Multiple Instance Learning (MIL) problem (Carbonneau et al., 2018; Song et al., 2019; Angelidis and Lapata, 2018): each conversation is a bag of window-level observations with a single bag-level label. We segment each task into 30-second windows, extract interactional and acoustic features, and estimate latent state trajectories using a Structured Temporal Projection (STP) that applies PCA to the observation sequence and recovers smooth latent trajectories. We compare this against static baselines and GRU models with mean-pool and gated attention pooling.

In this work, our primary contributions are:

1. A formal framing of dyadic collaborative conversation as a partially observable dynamic system with weak supervision, connecting classical DST to human–human conversation analysis and the MIL literature.

2. An empirical comparison of six models—static baselines, GRU with mean and attention pooling, and temporal models—on a total of 461 dyad-task sequences from AVCAffe dataset (Sarkar et al., 2023) under strict leave-one-dyad-out cross-validation.

3. Evidence that gated attention-based MIL pooling (Ilse et al., 2018) substantially recovers GRU performance (ΔCCC $+0.14$–0.17 over mean pooling), but with high fold-level variance, revealing that learned attention overfits to dyad-specific patterns at this dataset scale.

4. A trajectory analysis that shows that temporal projection reveals latent state evolution—escalation, convergence, stability—that vary systematically by task type and cognitive demand, patterns that are invisible to aggregate summaries.

2 Related Work

Dialogue state tracking. Classical DST models dialogue as a belief state over discrete slot-value pairs, updated at each turn (Young et al., 2013; Williams and Young, 2007). The POMDP dialogue manager (Young et al., 2010) provides a formal probabilistic foundation for partially observable dialogue state for modeling the inherent uncertainty about the unobservable states, such as cognitive load. Recent neural approaches focus on task-oriented DST (Jacqmin et al., 2022), but remain tied to semantic, explicitly annotated states. Heck et al. (2022) demonstrate that DST can be learned with weak supervision and sparse labels, a setting we extend to social-cognitive state in human–human conversation. However, prior work has considered dialogue states as discrete variables. Our work advances this line by treating conversational state as a latent *continuous* variable tracked over time, without turn-level annotations.

Social interaction dynamics. Vinciarelli et al. (Vinciarelli et al., 2009) established social signal processing as the study of behavioral signals encoding social and affective states in human interaction. Pickering and Garrod (Pickering and Garrod, 2004) show that interlocutors in a conversation align at multiple levels through interaction. Levinson and Torreira (Levinson and Torreira, 2015) characterizes the temporal organization of turn transitions. Interaction timing features—overlap, response latency, silence, speech share—are central to our observable feature set, grounded in the theoretical claim that these signals reflect a latent coordination state.

Cognitive load from speech behaviors. Acoustic and behavioral correlates of cognitive load have been studied in aviation (Yang et al., 2023), driving (Vukovic et al., 2021), and related single-speaker settings. In dyadic collaborative settings, task-level cognitive load varies substantially between speakers (A–B correlation $r = 0.30$–0.34 in our data), motivating a dyad-level modeling approach to capture the dyad-specific variability. To

the best of our knowledge, no prior work has explored how cognitive load *evolves* within a task in dyadic conversation, rather than describing the evolution of cognitive load as a static post-task outcome, which we address in this work.

MIL in dialogue and sequential data. Multiple Instance Learning has been applied to dialogue for satisfaction analysis (Song et al., 2019; Carbonneau et al., 2018), where conversational turns form a bag and utterance-level labels are unavailable, a structure analogous to ours. Our application extends MIL to dyadic social-cognitive state, using 30-second windows as instances and task-level NASA-TLX labels as the bag-level signal as weak supervision. As mean pooling considers each window to be equally informative within a task-level conversation, Ilse et al.'s gated attention pooling (Ilse et al., 2018) provides a natural replacement for mean pooling in this setting, learning which windows are most predictive of the bag label.

Weak supervision and temporal modeling. State-space and Kalman-filter models have been applied to prosody modeling, affect tracking (Nicolaou et al., 2011), and dialogue management (Williams and Young, 2007; Shumway and Stoffer, 2000). We distinguish two approaches to trajectory estimation used in this work. The first, our Principal Component Analysis + temporal smoothing, applies a static linear projection to the observation sequence and smooths the result with a Savitzky-Golay filter (Savitzky and Golay, 1964). Note that it does not model temporal dynamics, and window order does not affect the projection matrix, rather indicating how the conversation moved through a static 2-D feature space, which is not an estimate of a latent dynamic state. The second approach, the Kalman + smoothing, fits a Linear Gaussian state-space model via per-sequence Expectation Maximization (Shumway and Stoffer, 1982) on training data, explicitly modeling how the latent state evolves from window to window through a learned transition matrix. In this way, this models the temporal structure of the latent state dynamics.

3 Problem Formulation

We model collaborative dyadic conversation as a partially observable temporal process. Unlike DST, which tracks semantic slot values, our goal is to infer *continuous social-cognitive conversational state* evolving over time, capturing properties related to interactional asymmetry, coordination, and cognitive difficulty.

Let a conversation instance i be segmented into a sequence of fixed temporal windows:

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

where each $x_t^{(i)} \in R^d$ is a feature vector from window t. We assume an unobserved latent state trajectory as,

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

where $z_t^{(i)}$ represents the latent conversational state at window t, intended to capture interactional asymmetry, coordination, and cognitive difficulty. These window-level states are *not* annotated. Supervision is available only at the task level:

$$y^{(i)} \in R,$$

such as NASA-TLX mental or temporal demand. The learning problem is therefore:

$$X^{(i)} \rightarrow Z^{(i)} \rightarrow y^{(i)}.$$

4 Dataset and Feature Construction

Structure of Multiple Instance Learning. Because labels are task-level while observations are window-level, each conversation is a *bag* of window observations with a single bag-level label—a weakly supervised MIL problem (Carbonneau et al., 2018). Unlike standard MIL, where the bag label reflects at least one positive instance, here the label reflects the aggregate cognitive experience across the task. This motivates temporal trajectory-to-label prediction rather than instance-level labeling as performed in most supervised learning tasks.

Primary targets. We treat *mental demand* and *temporal demand* as primary supervision signals. These align closely with temporally organized conversational behavior: turn-taking speed, interruptions, overlaps, silences, and participation imbalance. They also show moderate inter-speaker agreement (A–B correlation $r = 0.34$ and $r = 0.30$ respectively). Based on this, we justify the decision of dyad-level aggregation and use the dyad mean as the target variable.

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 $\times$ 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.

Weak supervision and interpretation. Because supervision is coarse, post-task, and subjective, we do not claim access to the true within-conversation cognitive state. Inferred latent states are model-based state representations evaluated through (1) conversation-level predictive performance and (2) qualitative trajectory analysis.

4.1 AVCAffe Dataset

We use AVCAffe (Sarkar et al., 2023), an audio-visual corpus of remote dyadic collaboration recorded over a videoconferencing platform. 106 participants from 18 countries completed up to nine collaborative tasks varying in cognitive demand: open discussion (7.5 min), sharing jokes (7.5 min), finding differences in pictures (10 min), map matching (2.5 min $\times$ 2), Lost at Sea decision-making (10 min), reading comprehension (5 min $\times$ 2), and a multi-task condition. We chose this dataset for several reasons: 1) the dataset is not collected from the internet, and 2) the speakers need to collaborate together to complete interactive tasks of varying structure (e.g., cognitive, collaborative, and open conversation). This task diversity makes AVCAffe well-suited for studying within-task state evolution and cross-task trajectory comparisons.

Per-task self-report labels were collected using the NASA Task Load Index (Hart and Staveland, 1988) (NASA-TLX) on a scale of 0–21. We focus on **mental demand** and **temporal demand** as primary targets, which showed the most consistent signal in exploratory analysis. Mental demand reflects the degree of mental and cognitive difficulty required by the task; temporal demand reflects perceived time pressure due to task urgency. The two dimensions are conceptually different, which is why they respond differently to task type. Both are rated on a 0–21 scale, where higher values indicate greater task load (Hart and Staveland, 1988).

Table 1 shows that temporal demand varies substantially by task type. This suggests that temporal demand separates task categories more distinctly: high-demand tasks score 14.7–15.8 on temporal demand versus 9.4–14.0 on mental demand, whereas low-demand tasks score 4.2–5.3 on temporal demand versus 5.1–7.2 on mental demand. The gap between low- and high-demand categories is approximately 10 points for temporal demand versus 5 points for mental demand. This is consistent with time pressure being more directly coupled with task structure and task objectives, while perceived mental difficulty varies more across dyads regardless of task type.

4.2 Window-Based Representation

We apply Silero VAD (Silero Team, 2024) to each speaker-separated audio stream to identify speech activity segments. Each task recording is divided into non-overlapping 30-second windows, yielding a mean of 8.7 windows per task (min: 2, max: 20). Windows where at least one speaker had insufficient speech activity were excluded, removing 1,576 windows (26.3% of 5,982 total). The remaining windows form ordered sequences $x_1, \ldots, x_T$ per dyad-task.

We note that Silero VAD does not distinguish speech from other vocal activity such as laughter or filled pauses. In tasks involving laughter between dyad partners (e.g., Sharing Jokes, Task 2), vocal overlap may not reflect conversational turn-taking

structure in the same way as speech overlap. We acknowledge this as a limitation and future work should apply a non-speech classification step to separate these non-speech signals for task types with affective potential.

4.3 Label Aggregation

For labels, both speakers completed separate NASA-TLX ratings after each task. We aggregate across participants by taking their mean to form a single dyad-level supervision target per task: $y = (y_A + y_B)/2$. This reflects the shared conversational context experienced by both speakers. The low A–B agreement (mental demand: $r = 0.34$, temporal demand: $r = 0.30$, mean $|y_A - y_B|$ exceeding 5 scale points) reflects that participants experience the same task differently, a form of measurement noise inherent to dyadic self-report that motivates the dyad-level weak supervision framing of our study. This aggregation is further motivated by the dyadic nature of the features themselves: interaction signals such as floor imbalance, overlap count, and turn count ratio are inherently joint features— they cannot be attributed to one speaker alone. A dyad-level supervision target is therefore consistent with dyad-level observations. We note that individual within-dyad experiences diverge, and predicting individual-level load from joint conversational signals is a harder and distinct problem that we leave as future work. Across 461 dyad-task sequences: mental demand mean $= 9.92 \pm 4.67$; temporal demand mean $= 9.11 \pm 5.67$ (scale 0–21).

4.4 Feature Families

Interaction features ($N = 29$) These features are extracted from the VAD-derived speech activity of both speakers, capturing turn-taking structure, floor control, coordination, and timing. Features include: talk time, speech share, floor imbalance, turn count, and duration, turn count ratio, overlap count, duration, and ratio, interruption counts, and imbalance, response latency statistics (mean, median, std, max), long-silence $1\ s - -3\ s$ count and duration, maximum inter-turn silence, and interruption rate per speaker. Response latency is undefined in windows with no speaker transitions (13.1% of windows) and is set to zero in those cases. We note that interaction features are selected because they capture *joint* dyadic behavior, and they cannot be computed from a single audio stream. We expect them to provide stronger cross-dyad generalization than acoustic features due to audio compression in remote conversation platforms and speaker-specific differences.

Acoustic relative features ($N = 5$) We use five dyad-relative features representing the difference between speakers (A minus B): pitch mean, pitch standard deviation, pitch range, voiced segments per minute, and voiced duration. Using relative rather than absolute features reduces sensitivity to the automatic gain control present in videoconferencing audio. We exclude absolute energy, and spectral features are excluded as they are unreliable under compression. The full list of features and their literature supporting their role in our study is added in the Appendix (Tables 4 and 5).

Feature Ablation Conditions In our experiments we consider three conditions: *interaction only* (29 features), *acoustic only* (5 features), and *interaction + acoustic* (34 features).

5 Models and Experimental Setup

5.1 Static Baselines Models

Ridge Regression. Each sequence is summarized by the mean-pooled feature vector $\bar{x} = \frac{1}{T}\sum_t x_t$. Ridge regression ($\alpha = 1.0$) predicts the task-level label. All temporal order and within-conversation variation are discarded.

Random Forest. Same mean-pooled input as Ridge is used, but modeled with a Random Forest regressor (300 trees, minimum leaf size 2).

5.2 Gated Recurrent Unit-based Models

With Mean Pooling. A single-layer Gated Recurrent Unit based neural network (Cho et al., 2014) with hidden dimension 64 processes the ordered window sequence. Mean pooling over valid hidden states produces a fixed-length embedding that is passed to a linear regression head (Trained for 30 epochs (Adam, lr$= 10^{-3}$), predictions averaged over 5 seeds). Mean pooling treats all windows as equally informative, a poor inductive bias for MIL supervision, where the load may be concentrated at specific moments instead of uniformly distributed over the entire dyad.

With Attention MIL Pooling. We replace mean pooling with the gated attention pooling mechanism as described by Ilse et al. (Ilse et al., 2018). For GRU hidden states $H = [h_1, \ldots, h_T]$, the attention weight per window is:

$$a_t = \frac{\exp\big(\mathbf{w}^\top(\tanh(\mathbf{V}h_t) \odot \sigma(\mathbf{U}h_t))\big)}{\sum_s \exp(\ldots)},$$

	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

Table 2: Cross-validation results across 53 dyads. Rows are grouped by model family: static, sequential neural, and temporal trajectory (proposed).

and the pooled representation is $z = \sum_t a_t h_t$. Padding positions are masked before the softmax predictions, and all other training settings are identical to the mean-pool setting.

5.3 Trajectory-based Models

Both trajectory-based models share a trajectory estimation step (unsupervised), followed by a Ridge regression head (supervised), but differ in whether the estimation step models temporal dynamics.

PCA + Temporal Smoothing. PCA (Pearson, 1901) projects from F=34 observed features to K=2 latent dimensions, fit on training sequences only. Critically, PCA is a static linear projection: it does not model temporal dependencies between windows, and the window order does not affect the projection matrix. The resulting 2D sequence $\hat{z}_1, \ldots, \hat{z}_T$ is then smoothed with a Savitzky-Golay filter (Savitzky and Golay, 1964) (window length 5, polynomial order 2) to suppress observation noise, producing a smooth trajectory. The two dimensions explain 21.0% and 15.6% of total observation variance (combined: 36.6%), providing an interpretable but partial low-dimensional representation of the observation sequence.

Kalman Smoother. The Kalman smoother fits a Linear Gaussian state-space model (LDS):

$$z_t = A\, z_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, Q) \tag{1}$$

$$x_t = C\, z_t + v_t, \quad v_t \sim \mathcal{N}(0, R) \tag{2}$$

where $z_t \in R^K$ is the latent state at window t, $A \in R^{K \times K}$ is the transition matrix capturing how state evolves over time, and $C \in R^{F \times K}$ is the emission matrix relating latent state to observations. Unlike PCA, the transition equation (1) makes z_t causally dependent on z_{t-1}, encoding an explicit model of temporal evolution.

Model Comparison. Parameters $\{A, C, Q, R, \mu_0, \Sigma_0\}$ are estimated via the Expectation-Maximization algorithm (Shumway and Stoffer, 1982), fit independently of each training sequence to avoid the masked array instability of batch EM at small dataset scales. Final parameter estimates are averaged across training sequences before applying the Rauch-Tung-Striebel smoother (Shumway and Stoffer, 2000) to recover the posterior trajectory $\hat{z}_1, \ldots, \hat{z}_T$ for each test sequence. When EM does not converge (fewer than 3 windows), the model falls back to PCA projection as the default.

Shared prediction. Both methods produce a 2D trajectory $\hat{z}_1, \ldots, \hat{z}_T$. From each trajectory, we extract 4 scalar statistics per latent dimension: temporal mean (overall state level), final value (end-of-task state), linear slope (temporal trend: positive = escalating, negative = converging), and temporal standard deviation (variability). A Ridge regression head then predicts task-level labels from these statistics.

5.4 Evaluation

We use Leave-One-Dyad-Out (LODO) cross-validation as an evaluation setup where all tasks from one dyad form the test set; remaining dyads form the training set. As each dyad participates in only one recording session, training and test sets across all folds have separate speakers. We report Concordance Correlation Coefficient (CCC) (Lin, 1989) as the primary metric (captures both correlation and scale agreement), with Mean Absolute Error (MAE) as a secondary metric. Results reported as mean $\pm$ std across 53 LODO folds for cross-validation and capture dyad variability.

6 Results

Table 2 reports prediction performance under LODO cross-validation using the 34 features set. Random Forest, operating on mean-pooled features, achieves the strongest predictive perfor-

mance (CCC $= 0.197$ for mental demand, 0.254 for temporal demand). The mean-pooling GRU does not improve over this static baseline (CCC ≈ 0.00), confirming that access to temporal order alone is insufficient when the pooling strategy is misaligned with the MIL supervision structure.

Replacing mean pooling with gated attention-based MIL pooling (Ilse et al., 2018) substantially recovers performance (CCC $= 0.141$ and 0.149 for mental and temporal demand, respectively). However, the high fold-level variance (± 0.258 and ± 0.231) reveals that learned attention weights do not generalize reliably across unseen dyads: individual folds range from CCC $= 0.64$ to CCC $= -0.42$, indicating dyad-specific attention patterns that fail to generalize under LODO evaluation.

For trajectory-based models, the two approaches reveal a task-dependent pattern. PCA + temporal smoothing, a static linear projection without modeling temporal dynamics, achieves CCC $= 0.129$ for mental demand and CCC $= 0.057$ for temporal demand. The Kalman smoother models latent state evolution through a learned transition matrix, achieves CCC $= 0.073$ for mental demand but CCC $= 0.125$ for temporal demand. This reversal is interpretable: temporal demand is driven by structured, time-pressured task dynamics (e.g., accumulating pressure in map-matching tasks) that the Kalman transition model can capture. On the other hand, mental demand is a more diffuse cognitive experience distributed across the conversation, where a static data-driven projection is more robust than parametric dynamics estimated from 53 dyads.

Both trajectory-based models offer substantially more stable generalization than the attention GRU (fold variance ± 0.111–0.154 vs. ± 0.231–0.258), motivating their use in small-dataset MIL regimes. The high standard deviations across all models (± 0.10–0.25) reflect substantial dyad-level heterogeneity as different dyads show substantially different load patterns, and any model averaging across this heterogeneity faces an inherent performance ceiling, as we observe here.

6.1 Feature Ablation

In Table 3, we compare three feature families using Random Forest. Dyad-relative acoustic features yield near-zero CCC for mental demand (0.044), confirming that speaker-relative pitch and voicing features carry minimal signal for cognitive workload in a videoconferencing setting. Interaction features are the primary contributor across both targets. Adding relative acoustic features provides a modest improvement for the temporal demand ($0.209 \rightarrow 0.254$), but negligible change for mental demand, consistent with the interpretation that temporal demand is driven by turn-taking pacing while mental demand reflects participation imbalance.

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)	$\mathbf{0.197} \pm .221$	$\mathbf{0.254} \pm .198$

Table 3: Feature ablation (Random Forest, CCC $\pm$ std across 53 LODO folds). Acoustic features = 5 dyad-relative pitch and voicing measures; interaction features = 29 turn-taking and timing measures.

7 Trajectory Analysis

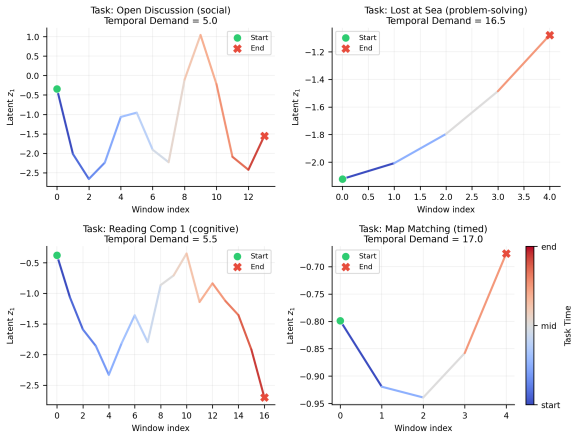
Beyond prediction accuracy, both trajectory-based models recover interpretable latent state trajectories $\hat{z}_1, \ldots, \hat{z}_T$ that static models cannot produce. We analyze these trajectories to understand *how* conversations evolve, not just where they end.

7.1 Trajectory examples

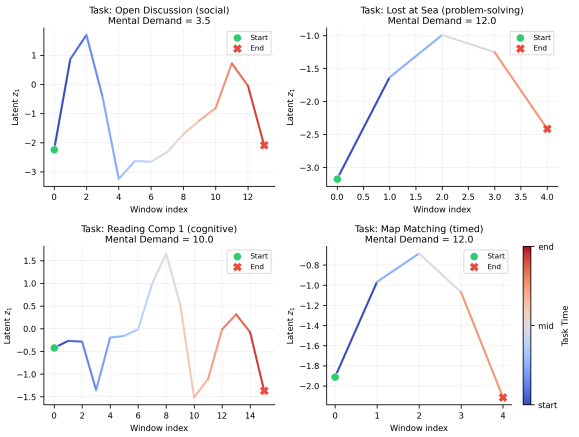
Figure 2 shows z_1 over window index for four representative dyad-tasks. High-demand tasks (Map Matching, Lost at Sea; temporal demand ≥ 16) show monotonically escalating z_1, consistent with accumulating time pressure. Low-demand tasks (Reading, open discussion) show flat or declining trajectories consistent with stable conversational dynamics. For mental demand, low-demand tasks follow the same trend as temporal, but high-demand tasks show higher-amplitude trajectories with greater variability.

7.2 Slope vs. Trajectory shape as a Latent Predictor

The Pearson correlation between the linear slope of z_1 and the task-level label across 450 dyad-task sequences is negligible for both mental demand ($r = -0.11$) and temporal demand ($r = -0.01$). Trajectory slope is therefore not a direct predictor of load under weak supervision—consistent with the constraint that the model optimizes observation structure, not label correlation, during trajectory estimation. Rather, trajectory *shape* provides analytical access to within-task dynamics that aggregate conversation summaries cannot expose. Figure 3



(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.

Figure 2: Latent state trajectories (z_1 over window index) for four representative task categories. Color encodes time within the task (blue = start, red = end); green dot = task start, red cross = task end.

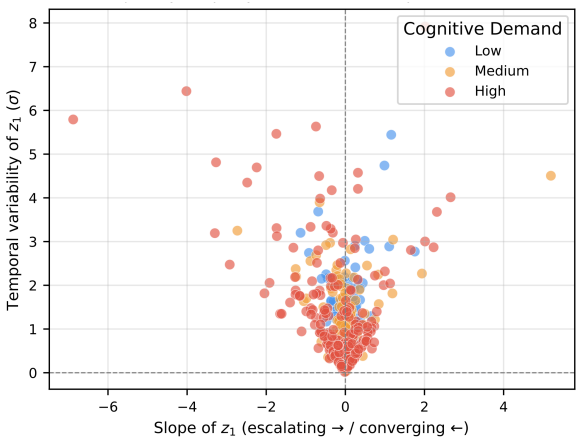


Figure 3: Trajectory shape scatter: linear slope of z_1 (temporal trend) vs. temporal variability (σ_{z_1}) per dyad-task, color-coded by task demand level (Low/Medium/High by task type). High-demand tasks (red) span the full slope range and are over-represented at high variability ($\sigma > 3$), indicating more dynamic and less predictable conversational state trajectories. Low-demand tasks cluster near the slope ≈ 0 with lower variability, reflecting stable dynamics.

confirms this pattern: high-demand tasks span the full slope range and are over-represented at high variability ($\sigma > 3$), while low-demand tasks cluster near slope ≈ 0.

The negligible slope-label correlation does not imply that trajectories are uninformative. Note that slope is one of eight statistics used jointly by the Ridge regression head; the trajectory's overall level, variability, and final value contribute alongside the slope value. More importantly, the trajectory figures demonstrate qualitatively distinct shapes across demand levels — monotonically escalating z_1 for high-demand tasks, flat or declining paths for low-demand tasks (Figure 2a) that are invisible to the static conversation mean regardless of slope's univariate correlation with the label.

8 Discussion

Pooling as architectural choice. The central empirical finding is the sharp contrast between the mean-pool GRU (CCC ≈ 0.00) and attention-pool GRU (CCC $= 0.14$–0.15). Mean pooling is misaligned in this problem, treating all windows as equally informative when the load may be concentrated at specific moments. Gated attention pooling (Ilse et al., 2018) recovers this issue, but high fold-level variance reveals that learned attention generalizes poorly under LODO with 53 dyads. Both trajectory-based models offer more stable generalization (fold variance ± 0.11–0.15), and produce interpretable trajectories that no pooling-based model can.

PCA vs. Kalman smoothing. The task-dependent reversal between the two trajectory methods is informative: the Kalman smoother outperforms PCA + smoothing for temporal demand (CCC 0.125 vs. 0.057), where task dynamics are structured and temporally ordered; PCA outperforms for mental demand (CCC 0.129 vs. 0.073), where the load is diffuse and static. A data-driven projection is more robust

than parametric dynamics estimated from 53 dyads, suggesting the value of modeling temporal evolution when the latent state is genuinely dynamic.

Future directions. An explicit cross-speaker coupling matrix for both speakers can formalize conversational dominance as asymmetry in state transition dynamics in future work. Once conversational state can be reliably tracked, a dialogue system can use the state estimate to trigger adaptive responses, connecting this to state-aware dialogue.

9 Conclusion

We proposed a weakly supervised framework for tracking latent conversational state in dyadic collaboration. Considering dyadic conversation as a partially observable dynamic system with MIL supervision, we compared six model classes on 461 dyad-task sequences from an audio-visual conversation corpus under strict leave-one-dyad-out cross-validation. Our results show that static interaction features are the strongest predictor for this problem, with high variance due to dyad-specific overfitting for pooling-based approaches. Trajectory estimation models based on PCA and Kalman with temporal smoothing provide more stable and interpretable latent trajectories, reflecting whether load dynamics are temporally structured or diffusely distributed. We also show that these trajectory shapes can be escalating, converging, or stable, as they vary systematically by task type and demand level in ways invisible to static summaries. These findings contribute a theoretically grounded step toward dynamic, interpretable representations of conversational state, with implications for dialogue modeling and state-aware collaborative support systems.

Limitations and Ethical Considerations

This work relies on a single dataset of 53 dyads recorded over a videoconferencing platform, which limits generalizability. Labels are collected post-task and may not reflect moment-to-moment cognitive state, which we address through our modeling choice. Linguistic features and visual signals were not included, which we plan to include in future work to understand the state space multimodally. Both trajectory models used here are linear; nonlinear extensions based on generative models are a next step for this work. The scope for temporal modeling is further constrained by task duration: short tasks (Map Matching 1 and 2 have 3.7 and 3.6 windows on average respectively; Table 1) provide limited temporal resolution for trajectory estimation component in our model. Trajectory-based models are most meaningful for longer tasks (Lost at Sea, 15.1 windows; Find Differences, 14.8 windows) that provide many observations, where within-task state evolution is observable across those many windows.

The key ethical concern for this work is demographic generalizability. Turn-taking norms, response latency, and overlap patterns vary systematically across languages and cultures (Stivers et al., 2009). A model trained on this particular dataset may not generalize equitably to speakers whose conversational norms differ from those in the training data distribution, and our evaluation does not assess performance disparities across speaker backgrounds. Future work should examine fairness across demographic groups before any deployment in real collaborative settings.

Acknowledgments

We thank the AIIM lab, Queen's University, Canada, for collecting, preparing, and making this dataset publicly available. We also thank the reviewers for their helpful comments to improve this work. This work was supported by the Henry Luce Foundation.

References

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, pages 103–111.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*, pages 699–708.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust dialogue state tracking with weak supervision and sparse data. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1175–1192.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2127–2136.

Luc Jacqmin, Lina M. Rojas-Barahona, and Benoit Favre. 2022. Revisiting state tracking for goal-oriented dialogue systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3140–3154.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731.

Lawrence I.-K. Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50(4):696–735.

Pritam Sarkar, Aaron Posen, and Ali Etemad. 2023. AVCAffe: A large scale audio-visual dataset of cognitive load and affect for remote work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 76–85.

Abraham Savitzky and Marcel J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Hugues Salamin, Anne Polychroniou, Fabio Valente, and Samuel Kim. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of Interspeech*, pages 148–152.

Robert H. Shumway and David S. Stoffer. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(5):253–264.

Robert H. Shumway and David S. Stoffer. 2000. *Time Series Analysis and Its Applications*. Springer, New York, NY.

Silero Team. 2024. Silero VAD: Pre-trained enterprise-grade voice activity detector. `https://github.com/snakers4/silero-vad`.

Kaisong Song, Lidong Bing, Wei Gao, Jun Lin, Lujun Zhao, Jiancheng Wang, Changlong Sun, Xiaozhong Liu, and Qiong Zhang. 2019. Using customer service dialogues for satisfaction analysis with context-assisted multiple instance learning. 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)*, pages 198–207, Hong Kong, China. Association for Computational Linguistics.

Tanya Stivers, Nicholas J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759. Visual and multimodal analysis of human spontaneous behaviour:.

Mihajlo Vukovic, Milan Stolar, and Margaret Lech. 2021. Cognitive load estimation from speech commands to simulated aircraft. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1011–1022.

Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Jianlong Yang, Hangyu Yang, Zhigang Wu, and Xuejun Wu. 2023. Cognitive load assessment of air traffic

controller based on SCNN-TransE network using speech data. *Aerospace*, 10(7):584.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialogue systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Appendix A: Feature Descriptions

Tables 4 and 5 list all 34 features used in the *interaction + acoustic* condition. Interaction features are computed from VAD-derived speech activity segments of both speakers jointly; acoustic relative features are computed per speaker and expressed as the difference A − B to capture asymmetry rather than absolute level.

#	Feature name	Definition	References
Speech share and floor control			
1	`talk_time_A`	Total voiced duration (s) for speaker A in the window	Vinciarelli et al. (2009)
2	`talk_time_B`	Total voiced duration (s) for speaker B in the window	Vinciarelli et al. (2009)
3	`speech_share_A`	Fraction of total speech time attributed to A: $\frac{tt_A}{tt_A+tt_B}$	Danescu-Niculescu-Mizil et al. (2012)
4	`speech_share_B`	Fraction of total speech time attributed to B	Danescu-Niculescu-Mizil et al. (2012)
5	`talk_time_ratio_A`	A's talk time as fraction of window duration	Vinciarelli et al. (2009)
6	`talk_time_diff_A_minus_B`	Signed difference in talk time: $tt_A - tt_B$ (s)	Danescu-Niculescu-Mizil et al. (2012)
7	`floor_imbalance`	Absolute talk time difference: $\lvert tt_A - tt_B \rvert$ (s)	Sacks et al. (1974)
Turn-taking structure			
8	`turn_count_A`	Number of complete turns produced by A	Sacks et al. (1974)
9	`turn_count_B`	Number of complete turns produced by B	Sacks et al. (1974)
10	`turn_count_ratio_A`	A's turns as fraction of total turns: $\frac{tc_A}{tc_A+tc_B}$	Gravano and Hirschberg (2011)
11	`mean_turn_duration_A`	Mean duration (s) per turn for A	Levinson and Torreira (2015)
12	`mean_turn_duration_B`	Mean duration (s) per turn for B	Levinson and Torreira (2015)
Simultaneous speech and overlap			
13	`overlap_count`	Number of windows with simultaneous speech from both speakers	Heldner and Edlund (2010)
14	`overlap_total_duration`	Cumulative duration (s) of simultaneous speech	Heldner and Edlund (2010)
15	`overlap_ratio`	Overlap duration as fraction of window duration	Gravano and Hirschberg (2011)
Interruptions			
16	`interruptions_A_to_B`	Count of A onset within B turn (A interrupts B)	Gravano and Hirschberg (2011)
17	`interruptions_B_to_A`	Count of B onset within A turn (B interrupts A)	Gravano and Hirschberg (2011)
18	`interruptions_total`	Total interruption count: $int_{AB} + int_{BA}$	Vinciarelli et al. (2009)
19	`interruption_imbalance_A_minus_B`	Signed interruption asymmetry: $int_{AB} - int_{BA}$	Danescu-Niculescu-Mizil et al. (2012)
20	`interruption_rate_A`	A's interruptions per minute of A speech	Gravano and Hirschberg (2011)
21	`interruption_rate_B`	B's interruptions per minute of B speech	Gravano and Hirschberg (2011)
Response latency			
22	`mean_response_latency`	Mean gap (s) between end of one turn and start of next (speaker change only)	Stivers et al. (2009); Levinson and Torreira (2015)
23	`median_response_latency`	Median inter-turn gap at speaker transitions	Levinson and Torreira (2015)
24	`std_response_latency`	Standard deviation of inter-turn gap at speaker transitions	Heldner and Edlund (2010)
25	`max_response_latency`	Maximum gap at speaker transition in the window	Heldner and Edlund (2010)
Silences			
26	`long_silence_count_ge_1p0s`	Count of inter-turn silences ≥ 1.0 s	Stivers et al. (2009)
27	`long_silence_total_duration_ge_1p0s`	Total duration (s) of silences ≥ 1.0 s	Heldner and Edlund (2010)
28	`long_silence_count_ge_3s`	Count of silences ≥ 3.0 s (extended pause)	Heldner and Edlund (2010)
29	`max_interturn_silence`	Maximum inter-turn silence (s) within the window	Stivers et al. (2009)

Table 4: Interaction features, computed from speech segments of both speakers within each 30-second window.

#	Feature name	Definition (A−B)	References
30	`pitch_mean_diff`	Mean fundamental frequency difference (Hz): $\bar{f}_{0A} - \bar{f}_{0B}$	Vukovic et al. (2021); Schuller et al. (2013)
31	`pitch_std_diff`	F0 standard deviation difference (Hz)	Yang et al. (2023)
32	`pitch_range_diff`	F0 range (max−min) difference (Hz)	Schuller et al. (2013)
33	`voiced_segments_per_min_diff`	Voiced segment rate difference (segments/min)	Vinciarelli et al. (2009)
34	`voiced_duration_diff`	Total voiced duration difference (s)	Vukovic et al. (2021)

Table 5: Acoustic relative features (5 total). Each feature is the signed difference A − B, capturing speaker asymmetry rather than absolute level. Absolute energy (RMS) and spectral features (centroid, bandwidth) are excluded as unreliable under videoconferencing automatic gain control. F0 is extracted via pYIN on voiced frames identified by Silero VAD (Silero Team, 2024).