Predicting Optimal Intervention Timing in MR Collaboration

Zhijie Liu
University of California, Irvine
Irvine CA US
zhijiel9@uci.edu

Yudong Wan
University of California, Irvine
Irvine CA US
yudongw4@uci.edu

Kai Yao
University of California, Irvine
Irvine CA US
kyao12@uci.edu

Yancheng Chen
University of California, Irvine
Irvine CA US
yanchec2@uci.edu

ABSTRACT

Mixed Reality (MR) environments offer valuable opportunities for collaborative learning, yet determining when teams require timely instructional intervention remains an open challenge. We propose an end-to-end framework that predicts optimal intervention timing from multimodal interaction logs. The approach integrates latent state discovery using a Gaussian Hidden Markov Model, a multi-method supervised feature selection process combining Random Forests, Mutual Information, LASSO, and RFE, and temporal prediction via sequence models including LSTM, GRU, and a lightweight Transformer. Using 32-second windowed features from eight training groups and evaluating on four unseen groups, the LSTM model achieves an F1 score of 80.23% and an accuracy of 79.46% for next-window intervention prediction. These findings demonstrate the feasibility of data-driven intervention timing and highlight the value of temporal modeling for supporting collaborative MR learning environments.

CCS CONCEPTS

· Computing methodologies → Machine learning → Learning paradigms; · Human-centered computing → Human computer interaction → Mixed reality;· Information systems → Data mining →Temporal data mining

KEYWORDS

Mixed Reality, Collaboration Analytics, Hidden Markov Models, Time Series Prediction, Deep Learning

ACM Reference format:

Zhijie Liu, Yudong Wan, Kai Yao and Yancheng Chen.2025.

Predicting Optimal Intervention Timing in MR Collaboration.

EECS215. *UCI, Irvine, CA, USA, 5 pages.*
https://github.com/LilyLiu945/MR-Collaboration-Intervention

1 Introduction

Collaborate through shared virtual content and collaborative multimodal interactions. However, assessing collaboration quality in mixed reality (MR) remains challenging because team behaviors are dynamic, latent, and not directly observable from raw interaction logs[1]. Educators and automated systems often lack timely metrics to reveal whether teams are functioning effectively or whether targeted support or interventions might be effective.

Most existing analysis methods rely on static features or manual annotation[2]. While these approaches provide rough summaries of overall performance, they overlook the temporal structure of collaboration. In practice, teamwork progresses through phases of initial exploration, collaborative problem-solving, and occasional disengagement[3]. Static representations fail to capture these transitions or predict how collaboration will evolve in the near future.

In this project, we utilize time-graph-based features extracted from mixed reality interaction logs to predict collaborative states in subsequent time windows. Our goal is to model collaboration as a dynamic process and develop a data-driven mechanism to predict shifts in team behavior. To achieve this, we propose a comprehensive workflow integrating unsupervised and supervised learning: (1) using Hidden Markov Models (HMM) to discover latent collaboration states; (2) employing multi-method feature selection to identify the most informative graph metrics; (3) Employing sequence models (LSTM, GRU, and Transformer encoders) to predict future collaboration states based on recent interaction windows.

Our contributions include:An end-to-end workflow for collaboration state discovery and prediction;HMM-based latent state inference;A multi-method feature selection framework; andComparative evaluation of LSTM, GRU, and Transformer models for temporal prediction performance[4].

2 Problem Definition

Understanding collaboration in mixed reality (MR) environments is difficult because group behavior unfolds dynamically, while multimodal interaction logs only provide indirect evidence of these changes. Although MR systems capture rich signals—such as spatial positioning, shared attention, conversational exchanges, and object interactions—these measurements do not directly reveal whether teams are collaborating effectively or experiencing disengagement or coordination issues. Consequently, instructors lack real-time, actionable indicators to decide when targeted intervention may enhance learning outcomes[5].

A major limitation of existing approaches is their reliance on static descriptors or retrospective manual annotation[6]. Such methods overlook how collaboration naturally transitions between phases of high engagement, low engagement, and occasional breakdowns. Moreover, collaborative states are rarely available as ground-truth labels, requiring them to be inferred from behavioral patterns rather than recorded explicitly.

To address this gap, we define the task as predicting the collaboration state of the next time window using the recent sequence of interaction features. Each window is represented by a graph-based vector summarizing participant interactions and engagement patterns in MR. The goal is to learn a model that maps several past windows to a predicted future state. These states are not predefined but are discovered using an unsupervised Hidden Markov Model, which provides discrete representations of latent collaboration quality[7].

This framing establishes two core requirements: collaboration must be modeled as a temporal process that depends on sequential context, and the underlying states are latent, requiring data-driven inference rather than direct observation. By treating collaboration prediction as supervised sequence classification over HMM-derived states, this approach enables fine-grained, forward-looking analysis of group behavior in MR learning environments[8].

3 Data Preparation

3.1 Window-Level Data Processing

We preprocess multimodal MR collaboration logs into standardized, model-ready window representations. This involves loading raw 32-second window data, restructuring modalities, aggregating node-level features, merging signals, handling missing values, and enforcing group-based splits.

Four datasets are used: windowed_metrics (dynamic multimodal features) and windowed_nodes (graph metrics) serve as primary inputs, while task_metrics and session_output are retained only for post-analysis. For each (group, window, modality), records are pivoted into a wide format. Node-level features—degree, betweenness, closeness, eigenvector centrality—are aggregated using mean and standard deviation to capture overall engagement and participation imbalance. The final dataset contains 54 dynamic features plus identifiers. Missing values are resolved via linear interpolation and zero-fill.

To prevent cross-team leakage, dataset partitioning is done strictly at the group level. Groups 1, 2, 4, 5, 9, 10, 11, and 12 form the training split, while groups 3, 6, 7, and 8 are held out for testing. A StandardScaler is fitted on training groups only and applied to all partitions. All processed splits, scalers, and feature lists are saved for reproducibility.

3.2 Feature Integration and Standardization

To ensure fair evaluation, all preprocessing respects the group-based train–test separation. Training groups (1, 2, 4, 5, 9, 10, 11, 12) produce 238 windows; test groups (3, 6, 7, 8) produce 197 windows, forcing models to generalize to unseen teams.

A StandardScaler is fit solely on training windows and then applied to validation and test sets to prevent leakage of statistical information. All intermediate outputs—including split feature matrices, fitted scalers, feature name files, and summary statistics—are stored for reproducible downstream experiments.

3.3 Unsupervised HMM Labeling

After standardization, each window becomes a 54-dimensional behavioral vector. These vectors serve as inputs to a Gaussian Hidden Markov Model for latent collaboration-state discovery. Identifiers (group, window index) are removed prior to training. Final matrices include 238 training windows and 197 testing windows.

A four-state Gaussian HMM with full covariance is trained using the Baum–Welch EM algorithm for up to 100 iterations, converging after five. The chosen configuration balances expressiveness and interpretability for continuous graph-based features. The HMM captures temporal evolution in collaboration and reveals latent behavioral modes.

Viterbi decoding generates a state per window. Training distributions are: State 0 (81), State 1 (34), State 2 (102), State 3 (21). In testing, States 0–2 appear, while State 3 does not. Feature means within each state indicate qualitative interpretations:

State 2: high collaboration (dense, cohesive interactions)

State 0: moderate collaboration

State 1: sparse communication

State 3: breakdown or imbalance

These discovered states guide subsequent label mapping.

3.4 Label Generation and Output

The four latent states are mapped into binary intervention labels: States 0 and 2 indicate normal collaboration (label 0), whereas States 1 and 3 represent low-quality or unstable interaction (label 1). After mapping, the training set contains 183 negative and 55 positive examples; the test set contains 174 negatives and 23 positives. These labels serve as ground truth for all supervised temporal models.

All labeling artifacts—including the trained HMM, decoded sequences, binary labels, and summary statistics—are saved to ensure consistency across modeling pipelines.

4 Feature Selection

To reduce redundancy in our 54-dimensional multimodal graph feature space and improve generalization, we adopt a supervised feature selection framework that integrates four complementary methods. Collaboration behaviors manifest through both linear and nonlinear relationships; therefore, relying on a single method would risk overlooking important predictive patterns. Our ensemble approach ensures that multiple aspects of feature relevance—statistical dependency, sparsity, interactions, and conditional importance—are jointly captured.

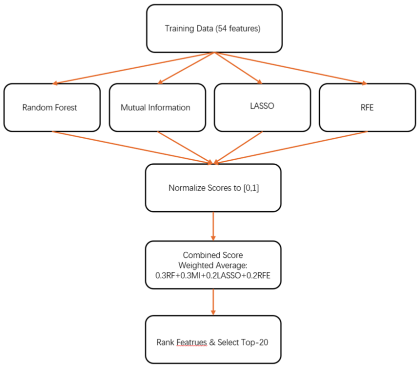
Training Data (54 features)
Random Forest
Mutual Information
LASSO
RFE
Normalize Scores to [0,1]
Combined Score
Weighted Average:
0.3RF+0.3MI+0.2LASSO+0.2RFE
Rank Featrues & Select Top-20

Figure 1. Overview of the supervised feature selection pipeline.

4.1 Feature Selection Method

Random Forest importance provides a non-linear measure of feature relevance by aggregating impurity reductions across trees, enabling the detection of interaction effects across conversation, proximity, and shared-attention modalities. Mutual Information evaluates the dependency between each feature and the intervention label without assuming linearity or monotonicity, making it suitable for complex collaboration dynamics. LASSO logistic regression introduces sparsity through L1 regularization and highlights interpretable linear predictors that strongly correlate with intervention needs. Finally, Recursive Feature Elimination (RFE) iteratively removes the least informative features under a logistic regression classifier, estimating feature importance in a conditional manner.

4.2 Ensemble Ranking and Feature Selection

Each method produces an independent importance ranking, which we normalize to the range [0,1] and combine via a weighted average to obtain a robust final score. This mitigates biases inherent to individual methods—for example, Random Forest's tendency to favor continuous features or LASSO's sensitivity to scaling. The top 20 features are selected based on the combined ranking, representing roughly 37% of the original feature space. These features primarily include centrality metrics, density and clustering measures, and reciprocity indicators, reflecting structural cohesion and participation balance in team interactions. Reducing the feature space improves interpretability, reduces overfitting risk, and enhances downstream sequence model performance.

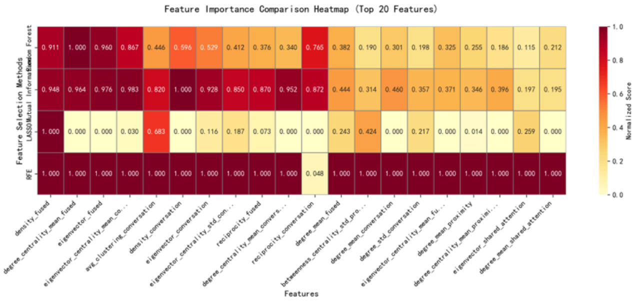

Figure 2. Normalized feature importance scores across the four selection methods.

5 Time-Series Prediction Models

To anticipate when an instructor intervention may be beneficial, we formulate the prediction task as a temporal sequence classification problem. Collaboration evolves as a continuous process, and transient behavioral changes often precede breakdowns. Therefore, incorporating temporal context is essential for accurate forecasting beyond static feature analysis.

5.1 Sequence Construction

We construct sequences using sliding windows: each training sample consists of three consecutive 32-second windows of selected features, used to predict the intervention label of the next window. This 3-window configuration balances temporal expressiveness with dataset size constraints, as longer sequences significantly reduce the number of available samples. To maintain behavioral coherence, sequences do not cross group boundaries. The final dataset includes 171 training, 43 validation, and 185 test sequences, each represented as a 3×20 tensor.

5.2 Model Architectures

We evaluate four model families: LSTM, GRU, a lightweight Transformer Encoder, and an MLP baseline. The recurrent models use stacked layers (64→32 units) to capture short- and mid-range dependencies. The Transformer employs positional embeddings and multi-head self-attention to model global relationships across the short input horizon. The MLP baseline flattens each sequence to test whether temporal modeling is necessary. Across models, parameter counts are controlled (35k–50k) to avoid overfitting.

5.3 Training and Model Selection

All models are trained with Adam (lr = 0.001), early stopping, and ReduceLROnPlateau scheduling. Weighted F1 serves as the primary evaluation metric due to class imbalance. LSTM and GRU achieve the best validation F1, but LSTM yields the highest AUC-ROC, indicating superior probabilistic separability. Therefore, LSTM is selected as the final model for intervention timing prediction.

Table 1 : Model Comparison

Model	Architecture	F1	Accuracy	AUC-ROC
LSTM	**2 layers (64→32) + dropout**	**0.9535**	**0.9535**	**0.9872**
GRU	2 layers (64→32)	0.9535	0.9535	0.9679
Transformer	Multi-head attention	0.9151	0.9070	0.9615
MLP	Feed-forward	< 0.85	< 0.85	-

6 Results & Discussion

6.1 Sequence Model Performance

On the held-out groups (3, 6, 7, 8), the LSTM achieves:

Table 2 : Test Result

Accuracy	Precision	Recall	F1 Score	AUC-ROC
0.7946	0.8105	0.7946	0.8023	0.6559

Performance decreases from validation due to distribution shift across teams, strong label imbalance, and limited temporal context (3-window sequences). These factors particularly affect minority states that require longer behavioral evidence to distinguish.

The confusion matrix $\begin{bmatrix}143 & 21\\ 17 & 4\end{bmatrix}$

shows strong detection of baseline states but difficulty identifying rare high-risk states (only 4 true positives). This reflects overlapping behavioral patterns between low-activity states and limited examples for the minority class. Additionally, many borderline episodes share structural similarity with normal collaboration phases, making them inherently ambiguous for short-window models.

Group-level evaluation further highlights heterogeneous collaboration styles: Group 7 shows highly stable behavior (F1 = 0.92), making patterns easier to model. Groups 3 and 6 exhibit noisier dynamics, suggesting inconsistent communication patterns or irregular task strategies. Group 8, despite having more windows, still achieves high accuracy, indicating that more data helps stabilize model learning. Overall, while minority recall remains a challenge, the model captures meaningful temporal signals that align with real group differences.

6.2 Intervention Effectiveness

We assess whether model-identified high-risk windows represent meaningful opportunities for timely assistance using a four-stage validation framework:

(1) identify high-risk windows,
(2) simulate counterfactual decay,
(3) apply intervention effects, and
(4) predict changes in team efficiency.

6.2.1 High-Risk Window Identification

The model flags 21/185 windows (11.4%), matching the ground-truth distribution and accurately reflecting group difficulty levels. This alignment suggests that the model captures not only short-term fluctuations but also broader team-level collaboration tendencies.

6.2.2 Feature-Level Collaboration Gains

Simulated interventions improve key collaboration metrics:

Table 3：Collaboration Improvement

Density	Reciprocity	Clustering	Eigenvector centrality	Avg. improvement
+0.1507	+0.1188	+0.1120	+0.0787	+0.45 SD

These increases show that predicted intervention timings tend to coincide with fragile yet recoverable collaboration phases. Because the intervention boosts structural connectivity and information flow within the network, even temporary improvements can alter the trajectory of subsequent windows. Moreover, improvements occur consistently across groups, suggesting that these high-risk windows share common structural signals detectable by the model.

6.2.3 Impact on Efficiency

Interventions lead to modest improvements in predicted efficiency:

- **0.17 sec** saved per intervention
- **~3.4 sec total**, ≈0.02% of task duration

Although the magnitude is small, this is consistent with the underlying transition dynamics.

Teams display a strong natural tendency to recover: (**P(1→0)=0.86**). This rapid recovery limits the additional benefit an external intervention can provide. Furthermore, the efficiency model is built on a small number of groups, which constrains its ability to capture complex nonlinear relationships between collaboration structure and task performance.

6.3 Overall Interpretation

- High-risk windows can be detected, though recall for rare states remains low.
- Interventions at predicted times substantially improve collaboration features.
- Efficiency gains exist but are small, constrained by rapid natural recovery and limited dataset size.

These findings show that temporal modeling provides useful insights for intervention timing, even if downstream efficiency improvements remain modest.

7 Conclusion & Future Work

This work demonstrates a complete pipeline for detecting collaboration breakdowns and estimating the potential benefit of timely interventions in MR-based teamwork. Sequence modeling captures meaningful temporal patterns, and simulated interventions consistently improve collaboration metrics.

However, efficiency improvements are limited by natural team dynamics, dataset size, and the rarity of high-risk states. Future work should focus on:

- Improving minority-class recall (e.g., focal loss, class-balanced training)
- Developing more robust efficiency models with larger datasets
- Learning realistic intervention effects from actual instructor logs
- A/B testing in real MR classrooms
- Incorporating richer multimodal signals (speech, emotion, gesture)

Despite limitations, our results provide evidence that model-driven intervention timing is feasible and promising for supporting collaborative MR learning environments.

REFERENCES

[1] Lehmann-Willenbrock, N., & Hung, H. (2024). A multimodal social signal processing approach to team interactions. *Organizational Research Methods*, *27*(3), 477-515.Lehmann-Willenbrock, N., & Hung, H. (2024). A multimodal social signal processing approach to team interactions. *Organizational Research Methods*, *27*(3), 477-515.

[2] Andrienko, Natalia, and Gennady Andrienko. "A visual analytics framework for spatio-temporal analysis and modelling." *Data Mining and Knowledge Discovery* 27.1 (2013): 55-83.

[3] Whittington, Colin. "A model of collaboration." *Collaboration in social work practice* (2003): 39-62.

[4] Zargar, S. "Introduction to sequence learning models: RNN, LSTM, GRU." *Department of Mechanical and Aerospace Engineering, North Carolina State University* 37988518 (2021).

[5] Olsen, Jennifer K., et al. "Temporal analysis of multimodal data to predict collaborative learning outcomes." *British Journal of Educational Technology* 51.5 (2020): 1527-1547.

[6] Staudt, Christian, et al. "Static and dynamic aspects of scientific collaboration networks." *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012.

[7] Xiao, Yue, et al. "Exploring latent states of problem‐solving competence using hidden Markov model on process data." *Journal of Computer Assisted Learning* 37.5 (2021): 1232-1247.

[8] Zheng, Yafeng, et al. "Investigating sequence patterns of collaborative problem-solving behavior in online collaborative discussion activity." *Sustainability* 12.20 (2020): 8522.