Note: Instead of missing windows, we set them to zero and continued to prevent loss of windows.

1 Cluster Characterization and Experiments

1.1 Cluster Characterization

For the robot-human interaction, we identified 5 distinct clusters after testing K = 2 through K = 10 clusters representing different communication strategies and adjusted them using intra-similarity.

1.2 Experiments

For the experiments the cluster predictions were extracted and compared against human annotated windows on the same test video data after training the model on 8 different videos, withholding 2.
Comparisons were made against human-annotations on the test data. To account for mismatches in the annotated data and the windows generated an algorithm was developed to be able to match the predictions to the windows based on the times given in the annotations and the end and start times of each window. The test data generated 100 windows to compare against the human annotations.
Based on these annotations and predictions the models accuracy is evaluated, applying a hierarchical rule-based label mapping.
To asses whether the model membership probabilities indicated the prediction reliability the maximum membership probability was extracted for each prediction. The margin between the top 2 cluster probabilities, calculating the entropy of the full distribution.
To test whether high confidence corresponds to a higher accuracy the confidence level was binned (0-30%, 30-50%, 50-70%) was created to asses the accuracy within each bin.
In order to identify systematic misclassifications patterns a confusion matrix is construct from the 100 test predictions.

The 5 clusters and there representations are shown below.

2 Results

This section will present the results obtained from the clusters' predictive performance against the annotated ground truth labels, including per-cluster accuracy metrics and probability-based analysis.

2.1 Overall Prediction Accuracy

The trained clustering model was evaluated over two held-out test sessions (111455, 114645) making 100 total prediction windows.

2.2 Per-Cluster Performance Metrics

The key observations from these metrics are:

- The **Intense Overlap** achieved the highest F1 Score of 51.69%, indicating that these features are the most distinctive.
- **Quiet Listening** shows that there is a high Recall of 66.67% but a low precision of 28.57%, indicating an over-prediction of this specific cluster.
- **Active Inquiry**, **Disengaged**, and **Collaborative Inquiry** achieved 0.00% accuracy, indicating that there are several false classifications with these clusters.

Table 1: Cluster Characterization and Communication Strategies

Cluster	Behavioral Profile	Communication Strategy	Interpretation
0: Active Inquiry	Elevated question frequency (wh_questions, yes_no_question); Low overlap ratio; Moderate speech activity without dominance; Smooth turn transitions.	Active Engagement; Smooth Turn Transition.	The students are actively exploring, asking questions, and engaging with the task while maintaining conversational order (no student talking over the other), indicating a focused manner without too much competition between the students.
1: Quiet Listening	High silence ratio; Low interruption rate; Reduced question frequency; Calming emotional prosody.	Passive Participation; Calming Prosody; Low Coordination.	The students are mainly observing the robot's actions, figuring out the next steps individually or reducing engagement among the students.
2: Intense Overlap	High overlap_ratio and overlap_duration; Elevated interruption_rate; Heightened emotional arousal (surprise, joy, or anger); Extended speaking turns.	Competitive Turn Claiming; Heightened Arousal Prosody; Extended Monologue.	Multiple students are competing to talk, often during moments of excitement, disagreement, or urgency, indicating a high-energy, potentially chaotic interaction phase.
3: Disengaged	Low collaboration_ratio; Minimal question activity; Absence of supportive backchanneling; Indicators of resistance or disagreement.	Resistance; Low Coordination.	The students show signs of withdrawal, disagreement, or task disconnection. This cluster may indicate confusion, frustration, or a loss of shared understanding between the students.
4: Collaborative Inquiry	High collaboration_ratio combined with high question_count; Elevated supportive backchanneling markers; Positive sentiment scores.	High Coordination; Active Engagement; Supportive Backchanneling.	This is the best state to see amongst students, indicating a large amount of problem solving, mutual support, and co-ordinated communication, representing the best student collaboration.

Table 2: Overall Prediction Metrics

Metric	Value
Total Test Samples	100
Overall Accuracy	37.00%
Correct Predictions	37
Incorrect Predictions	63

Table 3: Per-File Accuracy Breakdown

Test Session	Acc.	Corr/Tot
111455_features	40.38%	21/52
114654_features	33.33%	16/48

Table 4: Per-Cluster Performance Metrics

Cluster	Precision	Recall	F1 Score	Predicted	Actual
Intense Overlap	53.49%	50.00%	51.69%	43	46
Quiet Listening	26.57%	66.67%	40.00%	49	21
Active Inquiry	0.00%	0.00%	0.00%	1	23
Disengaged	0.00%	0.00%	0.00%	2	6
Collaborative Inquiry	0.00%	0.00%	0.00%	5	4

2.3 Confusion Analysis

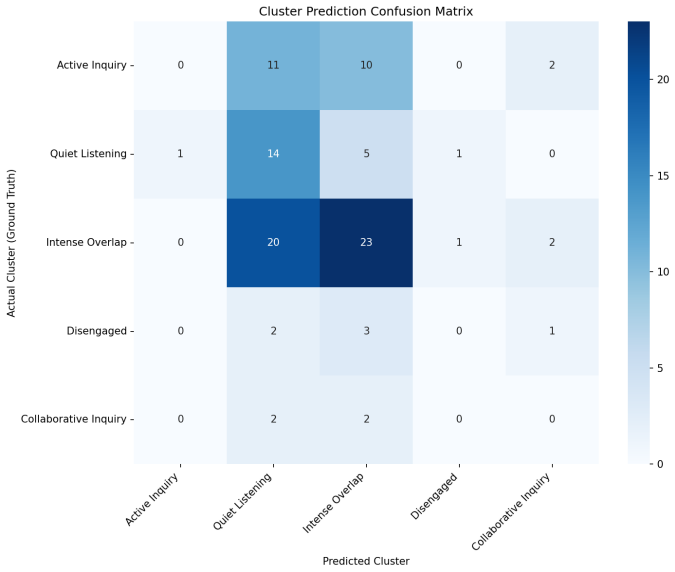

Figure 1: Cluster Prediction Confusion Matrix

The confusion matrix reveals and indicates that there are several misclassifications patterns. The most common misclassifications that can be seen are:

The model exhibits a strong bias towards predicting Quiet Listening and Intense Overlap, which together account for 92% of the total predictions.

2.4 Confidence & Calibration Analysis

From the above results, we can see that **Quiet Listening** shows the highest average confidence level despite low accuracy (overconfidence), while **Intense Overlap** is underconfident.

Table 5: Common Misclassifications

Ground Truth	Predicted As	Count
Intense Overlap	Quiet Listening	20
Active Inquiry	Quiet Listening	11
Active Inquiry	Intense Overlap	10
Quiet Listening	Intense Overlap	5
Disengaged	Intense Overlap	3

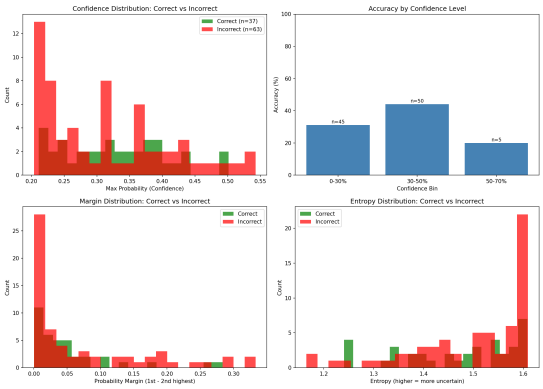

Figure 2: Confidence Distributions

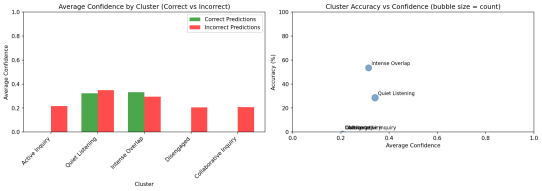

Figure 3: Accuracy vs Confidence

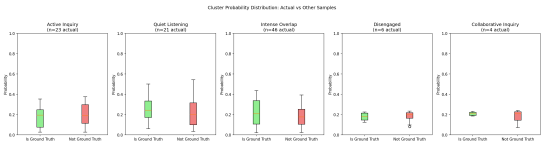

Figure 4: Cluster Probabilities (Boxplots)

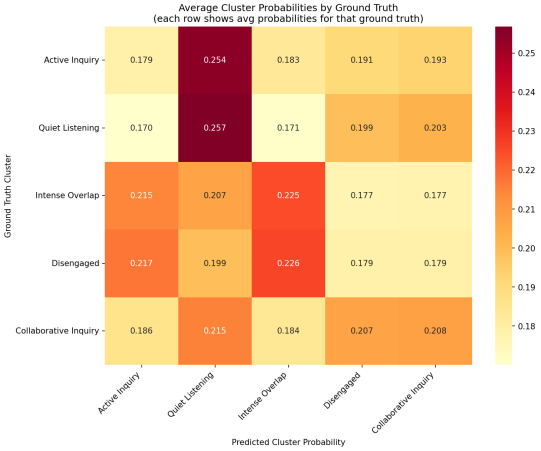

Figure 5: Probability Distribution (Heatmap)

Table 6: Cluster Confidence and Entropy Statistics

Cluster	Count	Acc (%)	Avg Conf	Margin	Entropy
Intense Overlap	43	53.5	0.315	0.021	1.486
Quiet Listening	49	28.6	0.342	0.116	1.473
Active Inquiry	1	0.0	0.217	0.003	1.606
Disengaged	2	0.0	0.204	0.003	1.609
Collab. Inquiry	5	0.0	0.208	0.001	1.609

3 Test Session 111455 Analysis

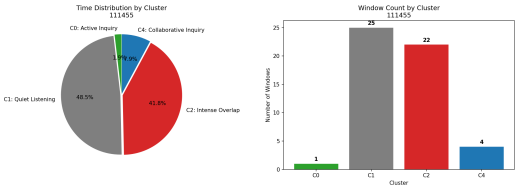

Figure 6: Cluster Distribution

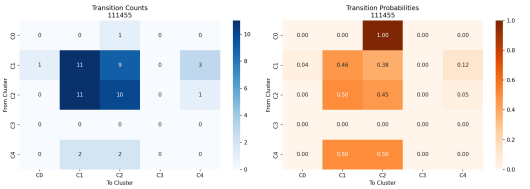

Figure 7: Transition Matrix

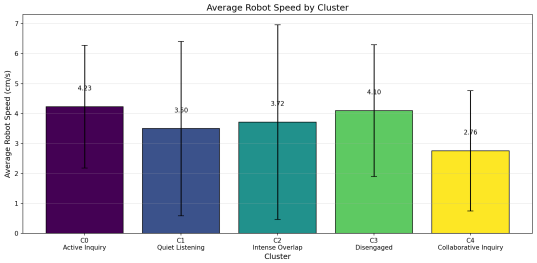

Figure 8: Robot Speed per Cluster

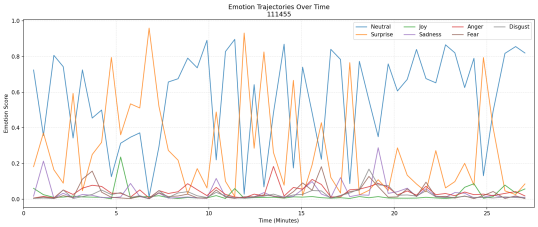

Figure 9: Emotion Trajectories

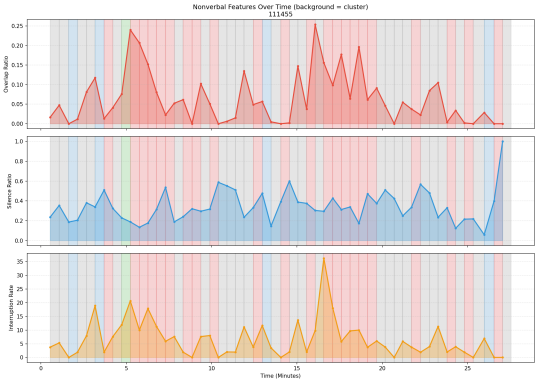

Figure 10: Nonverbal Features

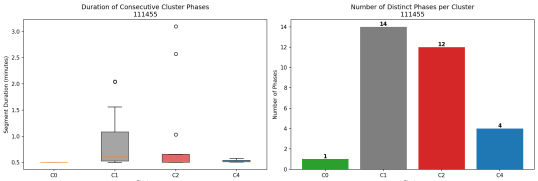

Figure 11: Cluster Duration Stats

The analysis of Session 111455 based on the features extracted from the video data helps reveal that:

- **Transitions:** Frequent transitions between Quiet Listening and Intense Overlap.
- **Robot Speed:** Higher speeds correlate with Active/Collaborative Inquiry.
- **Emotion:** Neutral emotion dominates; arousal spikes during Intense Overlap.
- **Nonverbal:** Overlap ratio peaks during Intense Overlap; silence peaks during Quiet Listening.
- **Duration:** Quiet listening phases are the longest (sustained passive observation).