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.
This small sample size creates a risk of sparse feature representation for rare classes, limiting the models ability to be able to learn robust decision boundaries for underrepresented communication states.
Comparisons were made against human-annotations on the test data. To compare the annotated data and the windows generated a many-to-one label aggregation strategy was implemented to be able to match the predictions to the time the start time from the excel file if it fell within the range of the JSON times. As there are 7 labels in the annotations and only 5 clusters a hierarchical rule-based mapping was used to match the predicted vs ground communication strategies. 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 based on the following rules and hierarchy:

- Intense Overlap: Competitive Turn Claiming OR Heightened Arousal OR Extended Overlap
- Disengaged: Disengaged OR Resistance present
- Collaborative Inquiry: High Coordination AND (Active Engagement OR Supportive Back channeling)
- Active Inquiry: Active Engagement OR Smooth Turn Transition (without High Coordination)
- Quiet Listening: Passive Participation OR Calming Prosody OR Low Coordination (default)

To asses whether the model membership probabilities indicated the prediction reliability, the maximum membership probability were extracted for each prediction. The margin between the top 2 cluster probabilities was used for 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%) to asses the accuracy within each bin.
In order to identify systematic misclassifications patterns a confusion matrix was construct from the 100 test predictions.

The 5 clusters and their respective 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.

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 back channeling markers; Positive sentiment scores.	High Coordination; Active Engagement; Supportive Back channeling.	This is the best state to see amongst students, indicating a large amount of problem solving, mutual support, and coordinated communication, representing the best student collaboration.

2.1 Overall Prediction Accuracy

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

Table 2: Overall Prediction Metrics: Hierarchical vs. Multi-Label

Metric	Hierarchical	Multi-Label
Total Test Samples	100	100
Overall Accuracy	34.00%	49.00%
Correct Predictions	34	49
Incorrect Predictions	66	51

Table 3: Per-File Accuracy Breakdown

Test Session	Hier. Acc.	Multi. Acc.	Hier. (C/T)	Multi. (C/T)
111455_features	38.5%	51.9%	20/52	27/52
114654_features	29.2%	45.8%	16/48	22/48

2.2 Per-Cluster Performance Metrics

Table 4: Per-Cluster Performance Metrics

Cluster	Precision	Recall	F1 Score	Predicted	Actual
Intense Overlap	54.29%	41.30.00%	46.91%	35	46
Quiet Listening	25.93%	66.67%	37.33%	54	21
Active Inquiry	20.00%	4.35%	7.14%	5	23
Disengaged	0.00%	0.00%	0.00%	5	6
Collaborative Inquiry	0.00%	0.00%	0.00%	1	4

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. As shown in Table 4, Active Inquiry had a Recall of 0.00% with only 1 predicted instance versus 23 actual ground truth samples the model essentially never predicted this class. Similarly, Disengaged (2 predicted vs 6 actual) and Collaborative Inquiry (5 predicted vs 4 actual) show severe under-prediction or misclassification.

2.3 Confusion Analysis

The confusion matrix 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.

Table 5: Multi-Label Per-Cluster Performance Metrics

Cluster	Precision	Recall	F1 Score	Predicted	Actual
Intense Overlap	54.29%	41.30%	46.91%	35	46
Quiet Listening	50.00%	69.23%	58.06%	54	39
Active Inquiry	0.00%	0.00%	0.00%	1	23
Disengaged	0.00%	0.00%	0.00%	5	16
Collaborative Inquiry	0.00%	0.00%	0.00%	1	4

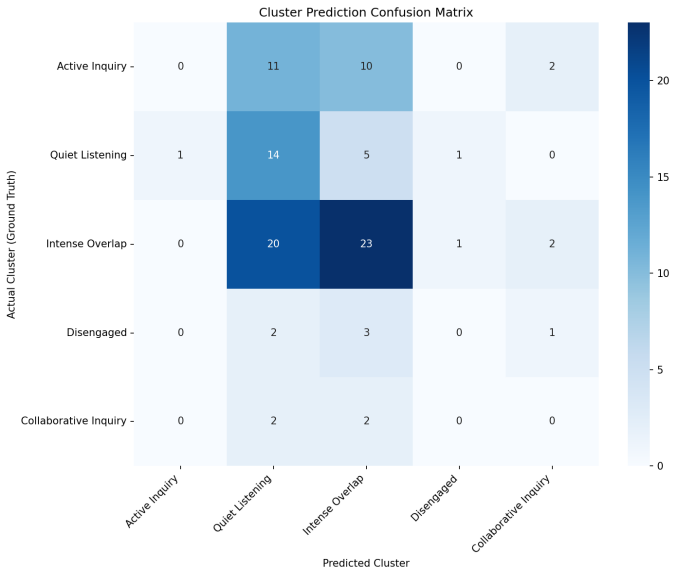

Figure 1: Cluster Prediction Confusion Matrix

Table 6: 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

Table 7: 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

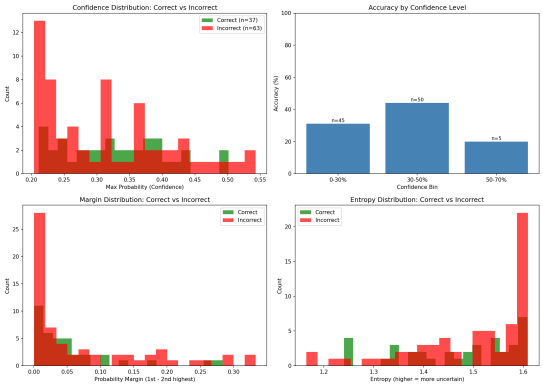

Figure 2: Confidence Distributions

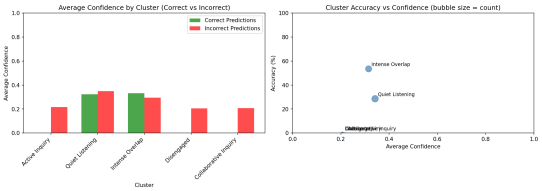

Figure 3: Accuracy vs Confidence

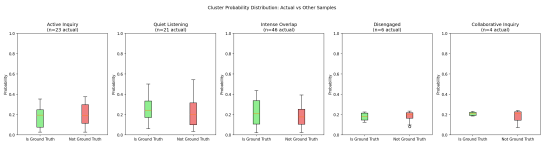

Figure 4: Cluster Probabilities (Boxplots)

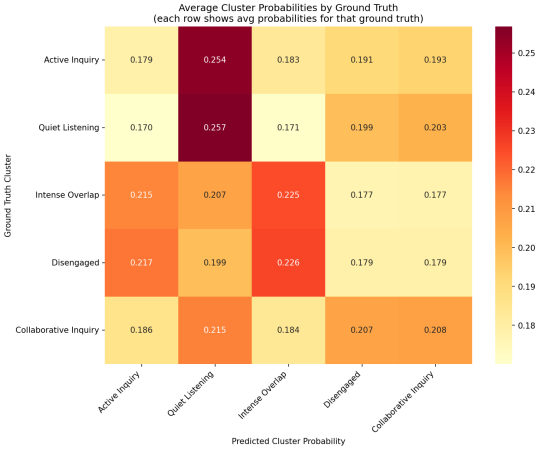

Figure 5: Probability Distribution (Heatmap)

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 under-confident.

3 Test Session 111455 Analysis

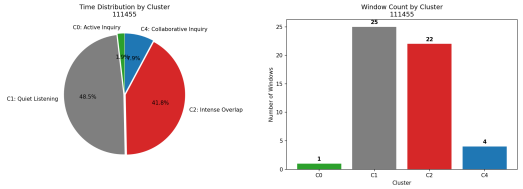

Figure 6: Cluster Distribution

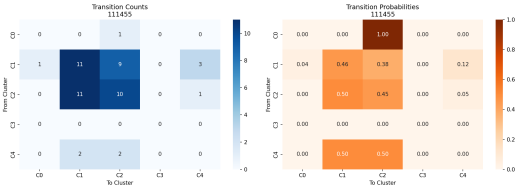

Figure 7: Transition Matrix

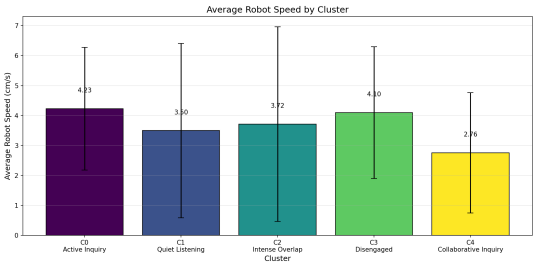

Figure 8: Robot Speed per Cluster

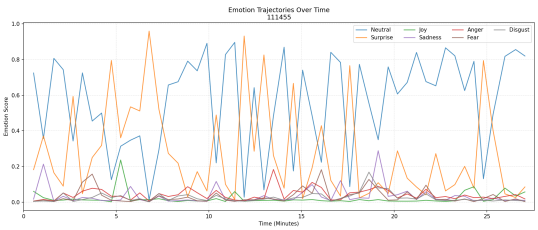

Figure 9: Emotion Trajectories

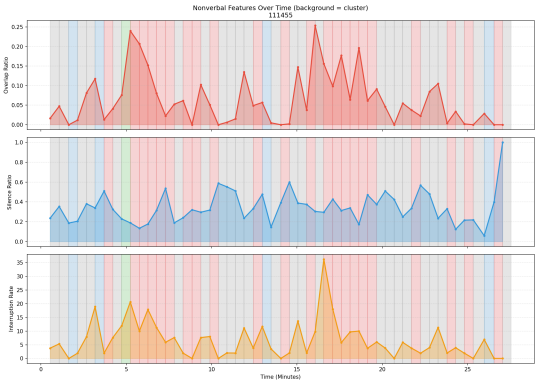

Figure 10: Nonverbal Features

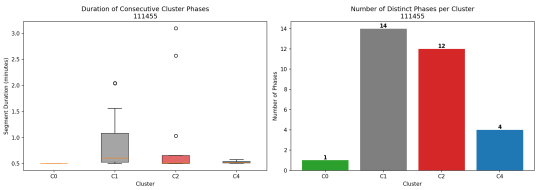

Figure 11: Cluster Duration Stats

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

- **Transitions:** Frequent transitions between Quiet Listening and Intense Overlap.
- **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).

The average robot speed per cluster was measured and obtained, however due to the accuracy of our model no conclusions can be made on whether speed has an impact on certain communication strategies such as Active Inquiry and Collaborative Inquiry as they were never accurately predicted.