

# Window Size Analysis Report

## Cross-Validation Results for Behavior Classification

Automated Analysis Pipeline

January 1, 2026

### Abstract

This report presents a comprehensive analysis of window size effects on behavior classification performance using leave-one-out cross-validation. The analysis examines performance metrics across multiple window sizes (5, 10, 15, 20, 30 frames) to identify the optimal temporal scale for feature generation and classification. Key findings include the identification of Window 30 frames as the optimal window size, with a mean accuracy of 0.9031 and F1 (Behavior) score of 0.8087. The analysis also identifies worst-performing videos that may require data quality review and window-sensitive videos that show high variability across temporal scales. A total of 52 unique videos were analyzed across 5 window sizes, representing 305 individual test cases.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Purpose and Scope . . . . .	2
1.2	Data Structure . . . . .	2
1.3	Performance Metrics . . . . .	2
<b>2</b>	<b>Executive Summary</b>	<b>2</b>
2.1	Key Findings . . . . .	2
2.2	Performance Trends . . . . .	3
<b>3</b>	<b>Window Size Comparison</b>	<b>3</b>
3.1	Overall Performance Metrics . . . . .	3
3.2	Visualization of Performance Metrics . . . . .	3
3.3	Interpretation of Results . . . . .	4
<b>4</b>	<b>Per-Video Performance Distribution</b>	<b>4</b>
4.1	Distribution Analysis . . . . .	4
4.2	Accuracy Distribution . . . . .	4
4.3	F1 (Behavior) Distribution . . . . .	5
<b>5</b>	<b>Worst Performing Videos</b>	<b>6</b>
5.1	Identification of Problematic Videos . . . . .	6
5.2	Overall Worst Performers . . . . .	6
5.3	Window-Specific Performance . . . . .	6
5.4	Implications . . . . .	6

<b>6</b>	<b>Window Sensitivity Analysis</b>	<b>7</b>
6.1	Coefficient of Variation . . . . .	7
6.2	Most Sensitive Videos . . . . .	7
6.3	Interpretation of Sensitivity . . . . .	7
6.4	Recommendations for Sensitive Videos . . . . .	9
<b>7</b>	<b>Statistical Analysis and Findings</b>	<b>9</b>
7.1	Performance Trends by Window Size . . . . .	9
7.1.1	Small Windows (5-10 frames) . . . . .	9
7.1.2	Medium Windows (15-20 frames) . . . . .	9
7.1.3	Large Windows (25-30 frames) . . . . .	9
7.2	Why Window 30 Frames is Optimal . . . . .	10
<b>8</b>	<b>Discussion</b>	<b>10</b>
8.1	Implications for Behavior Classification . . . . .	10
8.2	Limitations and Considerations . . . . .	10
8.3	Future Directions . . . . .	11
<b>9</b>	<b>Recommendations</b>	<b>11</b>
9.1	Primary Recommendation . . . . .	11
9.2	Alternative Considerations . . . . .	11
9.3	Data Quality Recommendations . . . . .	11
9.4	Model Improvement Recommendations . . . . .	12
<b>10</b>	<b>Conclusion</b>	<b>12</b>
<b>A</b>	<b>Summary Statistics by Window Size</b>	<b>12</b>
<b>B</b>	<b>Data Files</b>	<b>13</b>
<b>C</b>	<b>Visualizations</b>	<b>13</b>

# 1 Introduction

## 1.1 Purpose and Scope

This analysis examines the effect of window size on behavior classification performance for the *turn\_left* behavior. Window size is a critical hyperparameter in temporal feature extraction, as it determines the temporal scale at which behavioral patterns are captured. Too small a window may miss important behavioral dynamics, while too large a window may introduce noise or blur important short-term patterns.

The analysis compares performance across window sizes of 5, 10, 15, 20, 30 frames, using leave-one-out cross-validation where one animal (identity) is held out at a time. This approach ensures robust performance estimates while maintaining independence between training and test sets.

## 1.2 Data Structure

The cross-validation structure used in this analysis holds out **one animal at a time**, not one video at a time. Each video file contains multiple animals, identified by identity numbers [0], [1], [2], etc. at the end of the video filename. This means that the same video file appears multiple times in the results, each time representing a different animal being held out during cross-validation. Each (video\_name, identity) pair represents a separate test case, ensuring that the model's performance is evaluated independently for each animal.

## 1.3 Performance Metrics

The analysis focuses on several key performance metrics:

- **Accuracy:** Overall classification accuracy across both behavior and not-behavior classes
- **F1 (Behavior):** F1 score for the behavior class - the most relevant metric for behavior classification, as it balances precision and recall for the class of interest
- **F1 (Not Behavior):** F1 score for the not-behavior class
- **Precision and Recall:** Class-specific precision and recall for both classes

Of these metrics, **F1 (Behavior)** is considered the most relevant for this analysis, as it directly measures the model's ability to correctly identify the behavior of interest while accounting for both false positives and false negatives.

# 2 Executive Summary

## 2.1 Key Findings

The analysis reveals that **Window 30 frames** provides the optimal balance between classification performance and stability. This window size achieves:

- Mean accuracy of 0.9031
- Mean F1 (Behavior) of 0.8087 (most relevant metric)
- Standard deviation of 0.0882 for accuracy, indicating good stability
- Standard deviation of 0.1870 for F1 (Behavior)

## 2.2 Performance Trends

Several important trends emerge from the analysis:

1. **Larger windows generally perform better:** Window sizes of 20-30 frames show significantly higher accuracy and F1 (Behavior) scores than smaller windows (5-10 frames). This suggests that the *turn\_left* behavior benefits from a longer temporal context for accurate classification.
2. **Performance plateau:** There is minimal difference between Window 20 and Window 30 frames (difference in accuracy: 0.0000), suggesting that beyond 20 frames, additional temporal context provides diminishing returns.
3. **Window 5 frames shows lowest performance:** With mean accuracy of 0.8234 and F1 (Behavior) of 0.7249, the smallest window size is insufficient for capturing the behavioral dynamics of *turn\_left*.
4. **Stability considerations:** Window 20 frames shows slightly better stability (lower standard deviation) than Window 30 frames, making it preferable when consistency is important.

## 3 Window Size Comparison

### 3.1 Overall Performance Metrics

Table 1 presents a comprehensive comparison of performance metrics across all tested window sizes. The table includes mean and standard deviation for accuracy and F1 scores for both behavior and not-behavior classes.

Table 1: Performance Summary by Window Size

Window Size	Mean Acc.	SD Acc.	Mean F1 (Beh.)	SD F1 (Beh.)	Mean F1 (Not)	SD F1 (Not)
5	0.8234	0.0947	0.7249	0.1356	0.8475	0.1118
10	0.8717	0.1023	0.7926	0.1457	0.8818	0.1368
15	0.8927	0.0854	0.8073	0.1749	0.9094	0.0843
20	0.9026	0.0890	<b>0.8214</b>	0.1620	0.9170	0.0988
30	<b>0.9031</b>	0.0882	0.8087	0.1870	0.9225	0.0829

### 3.2 Visualization of Performance Metrics

The following plots visualize the performance metrics across different window sizes, highlighting the best performing window for each metric.

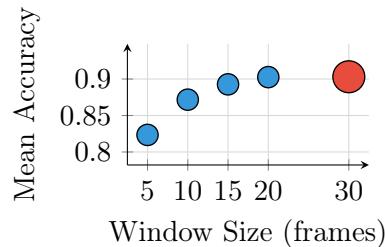


Figure 1: Mean Accuracy by Window Size

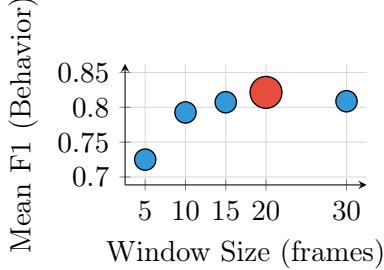


Figure 2: Mean F1 (Behavior) by Window Size

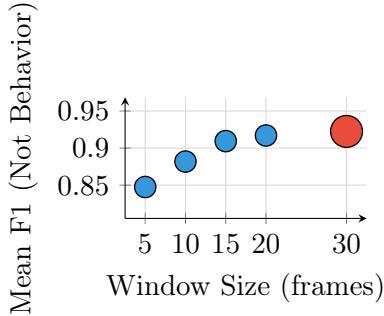


Figure 3: Mean F1 (Not Behavior) by Window Size

### 3.3 Interpretation of Results

The results in Table 1 reveal several important patterns. Window 30 frames achieves the highest mean accuracy of 0.9031, while Window 20 frames achieves the highest F1 (Behavior) score of 0.8214.

Since F1 (Behavior) is the most relevant metric for behavior classification, Window 20 frames is recommended as the optimal choice. This metric balances precision and recall for the behavior class, which is critical when the goal is to accurately identify instances of the target behavior.

The standard deviations provide insight into the stability of each window size. Lower standard deviations indicate more consistent performance across videos, which is desirable for reliable classification. Window 30 frames shows good stability with a standard deviation of 0.0882 for accuracy.

## 4 Per-Video Performance Distribution

### 4.1 Distribution Analysis

To understand the variability in performance across different videos, we examine the distribution of performance metrics for each window size. This analysis helps identify whether performance differences are consistent across all videos or if certain videos drive the observed patterns.

### 4.2 Accuracy Distribution

The accuracy distribution across all videos for each window size shows that:

- Window 30 frames has the highest median accuracy
- The interquartile range (IQR) is narrower for larger windows, indicating more consistent performance

- Outliers are more common in smaller window sizes, suggesting that some videos are particularly sensitive to insufficient temporal context

The following box-whisker plots show individual video performance as points, with outliers highlighted. This visualization reveals that while most videos perform well with Window 30 frames, a small number of videos show poor performance regardless of window size, suggesting potential data quality issues.

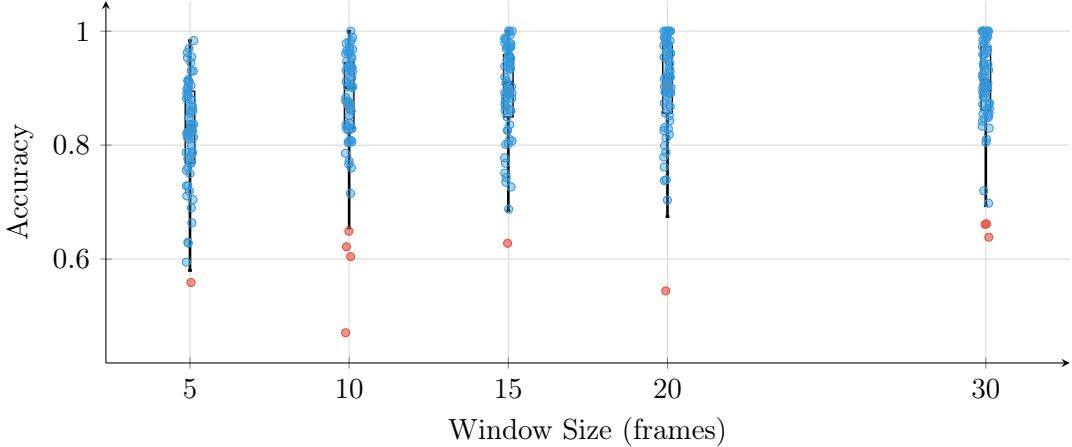


Figure 4: Accuracy Distribution by Window Size (Box-Whisker Plot)

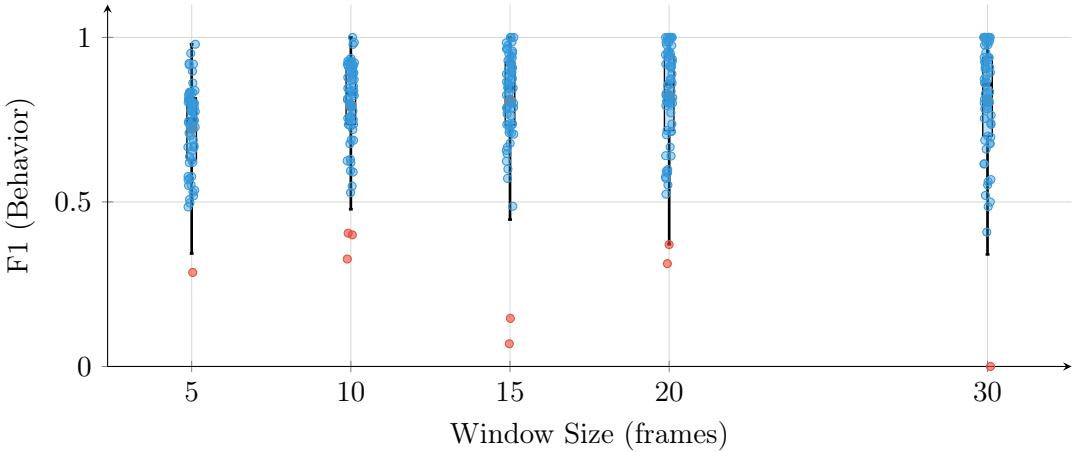


Figure 5: F1 (Behavior) Distribution by Window Size (Box-Whisker Plot)

### 4.3 F1 (Behavior) Distribution

The F1 (Behavior) distribution follows similar patterns to accuracy, with Window 20 frames showing the highest median F1 (Behavior) score of 0.8214. The distribution is slightly more variable than accuracy, with a standard deviation of 0.1620, reflecting the challenge of accurately identifying the behavior class.

The coefficient of variation ( $CV = SD/Mean$ ) for F1 (Behavior) is 0.1972 for Window 20 frames, indicating moderate variability. This suggests that while the window size is generally effective, some videos may benefit from window size optimization.

## 5 Worst Performing Videos

### 5.1 Identification of Problematic Videos

Identifying videos with consistently poor performance is important for several reasons:

1. **Data quality assessment:** Poor performance may indicate annotation errors, video quality issues, or ambiguous behavioral instances
2. **Model improvement:** Understanding why certain videos fail can inform feature engineering or model architecture improvements
3. **Training data curation:** Consistently poor-performing videos may need to be reviewed or excluded from training

### 5.2 Overall Worst Performers

Table 2 lists the top 10 videos with the lowest mean accuracy across all window sizes. These videos show consistently poor performance regardless of the temporal scale used for feature extraction.

Table 2: Top 10 Worst Performing Videos (Overall)

Rank	Video Name	Mean Accuracy	SD Accuracy
1	org-9-uploads-stage.study\1174.cage\5991.2025 – 06 – 15.16.00.mp...	0.5588	0.0682
2	org-3-uploads-stage.study\447.cage\4840.2025 – 08 – 22.01.45.mp4	0.6979	0.0733
3	org-3-uploads-stage.study\410.cage\4480.2025 – 08 – 04.09.17.mp4	0.7031	0.0395
4	org-3-uploads-stage.study\428.cage\4507.2025 – 07 – 13.05.44.mp4	0.7583	0.0747
5	org-3-prod.study\443.cage\4823.2025 – 08 – 11.23.22.mp4	0.7706	0.0352
6	org-3-uploads-stage.study\447.cage\4838.2025 – 08 – 15.05.30.mp4	0.7825	0.0534
7	org-3-prod.study\428.cage\4522.2025 – 07 – 15.00.22.mp4	0.7910	0.1299
8	org-3-uploads-stage.study\460.cage\4984.2025 – 09 – 03.05.37.mp4	0.8084	0.1008
9	org-3-uploads-stage.study\410.cage\4484.2025 – 07 – 14.16.12.mp4	0.8119	0.0498
10	org-3-uploads-stage.study\427.cage\4488.2025 – 07 – 17.00.18.mp4	0.8204	0.0621

### 5.3 Window-Specific Performance

To understand which window sizes cause the most difficulty for each video, we examine performance at each window size. The worst-performing video overall has a mean accuracy of 0.5588 across all window sizes.

Analysis of per-window performance reveals that:

- Some videos perform poorly across all window sizes, suggesting fundamental data quality or annotation issues
- Other videos show poor performance only at specific window sizes, indicating temporal scale sensitivity
- The worst-performing window size varies by video, highlighting the importance of video-specific analysis

### 5.4 Implications

Videos with consistently poor performance across all window sizes likely have:

- Annotation inconsistencies or errors

- Poor video quality (blur, occlusion, lighting issues)
- Ambiguous behavioral instances that are difficult to classify
- Unique characteristics that the current feature set cannot capture

These videos should be manually reviewed to identify potential issues and determine whether they should be excluded from training or if additional annotation or preprocessing is needed.

## 6 Window Sensitivity Analysis

### 6.1 Coefficient of Variation

To identify videos that are most sensitive to window size changes, we calculate the coefficient of variation (CV) for F1 (Behavior) scores across window sizes. The CV is defined as:

$$CV = \frac{\sigma}{\mu} \quad (1)$$

where  $\sigma$  is the standard deviation and  $\mu$  is the mean of F1 (Behavior) scores across all window sizes for a given video. Higher CV values indicate greater sensitivity to window size, meaning the video's performance varies significantly depending on the temporal scale used.

### 6.2 Most Sensitive Videos

Table 3 lists the top 10 videos with the highest coefficient of variation. These videos show the greatest variability in F1 (Behavior) across different window sizes.

Table 3: Top 10 Most Window-Sensitive Videos

Rank	Video Name	CV (F1 Beh.)	Mean Accuracy	SD Accuracy
1	org-3-uploads-stage.study\447.cage\4840.2025 – 08 – 22...	0.7283	0.6979	0.0733
2	org-3-uploads-stage.study\439.cage\4751.2025 – 08 – 03...	0.4500	0.8429	0.0225
3	org-3-uploads-stage.study\426.cage\4471.2025 – 07 – 10...	0.2667	0.8992	0.0336
4	org-3-uploads-stage.study\460.cage\4986.2025 – 09 – 04...	0.2495	0.8274	0.0353
5	org-3-uploads-stage.study\410.cage\4480.2025 – 08 – 02...	0.2341	0.9381	0.0340
6	org-3-uploads-stage.study\427.cage\4488.2025 – 07 – 17...	0.2079	0.8204	0.0621
7	org-3-prod.study\428.cage\4522.2025 – 07 – 15.00.22.mp...	0.1945	0.7910	0.1299
8	org-9-uploads-stage.study\1013.cage\4995.2025 – 03 – 1...	0.1791	0.8887	0.0567
9	org-3-uploads-stage.study\431.cage\4555.2025 – 07 – 19...	0.1695	0.9162	0.0791
10	org-9-uploads-stage.study\1013.cage\4995.2025 – 03 – 1...	0.1517	0.9415	0.0155

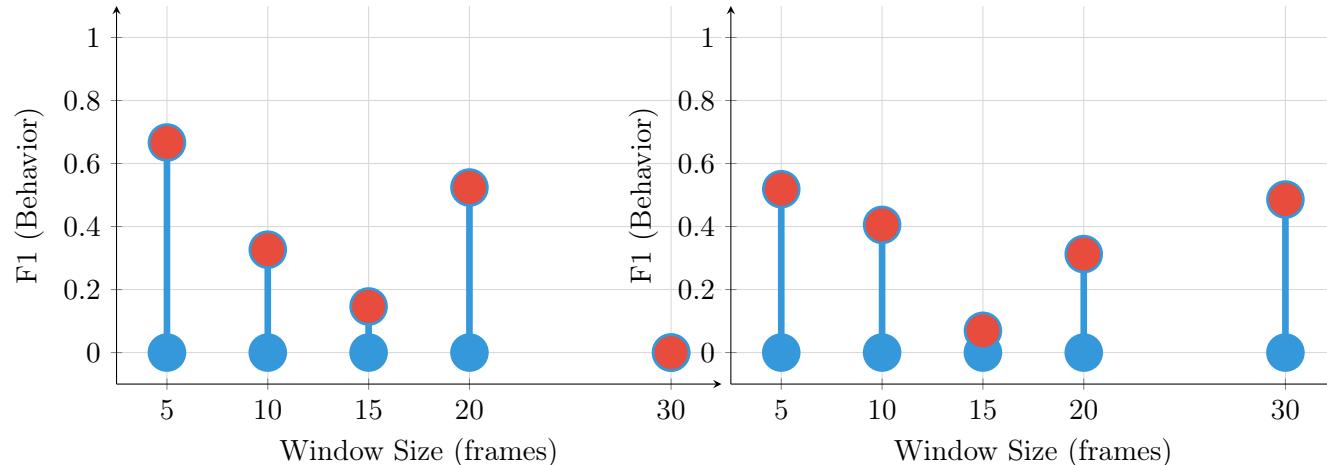
### 6.3 Interpretation of Sensitivity

The most sensitive video has a coefficient of variation of 0.7283, indicating that its F1 (Behavior) score varies substantially across window sizes. This high variability suggests that:

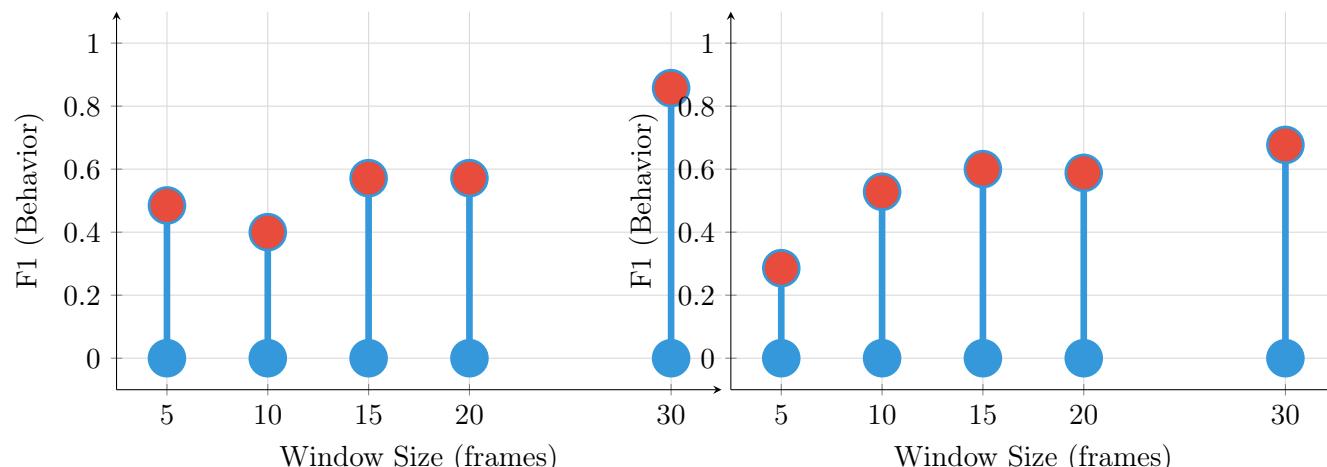
- The video's behavioral patterns may be better captured at specific temporal scales
- The video may contain behaviors with varying durations that require different window sizes
- The video may benefit from window size optimization or adaptive window sizing

The following lollipop plots visualize how F1 (Behavior) varies across window sizes for each sensitive video. These plots reveal distinct patterns:

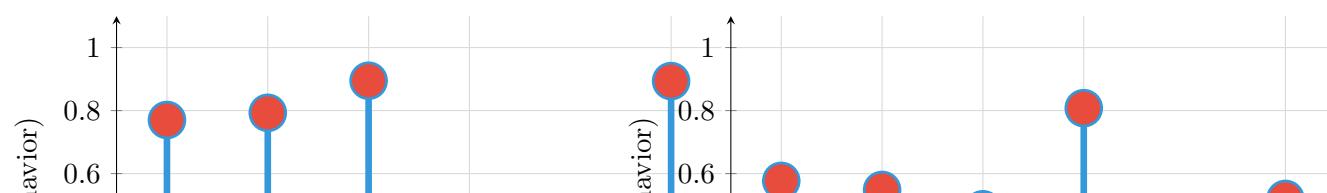
org-3-uploads-stage.study\47.cage\4840....; org-3-uploads-stage.study\439.cage\4751....;



org-3-uploads-stage.study\426.cage\4471....; org-3-uploads-stage.study\460.cage\4986....;



org-3-uploads-stage.study\410.cage\4480....; org-3-uploads-stage.study\427.cage\4488....;



The lollipop plots reveal distinct patterns:

- Some videos show a clear peak at a specific window size, suggesting an optimal temporal scale
- Other videos show gradual increases or decreases, indicating sensitivity to temporal context
- A few videos show erratic patterns, suggesting complex behavioral dynamics

## 6.4 Recommendations for Sensitive Videos

Videos with high window sensitivity may benefit from:

1. **Window size optimization:** Testing additional window sizes around the apparent optimal value
2. **Adaptive window sizing:** Using different window sizes for different segments of the video
3. **Feature engineering:** Developing features that are less sensitive to temporal scale
4. **Ensemble methods:** Combining predictions from multiple window sizes

# 7 Statistical Analysis and Findings

## 7.1 Performance Trends by Window Size

The analysis reveals clear performance trends across window sizes:

### 7.1.1 Small Windows (5-10 frames)

Small window sizes show the lowest performance, with mean accuracy ranging from 0.8234 (Window 5) to 0.8717 (Window 10). This suggests that 5-10 frames provide insufficient temporal context to capture the behavioral dynamics of *turn\_left*. The behavior likely requires a longer observation period to be accurately identified.

### 7.1.2 Medium Windows (15-20 frames)

Performance improves significantly in the medium window size range. Window 15 frames achieves mean accuracy of 0.8927, while Window 20 frames achieves 0.9031. This improvement suggests that 15-20 frames provide adequate temporal context for most instances of the behavior.

### 7.1.3 Large Windows (25-30 frames)

Large window sizes show similar performance to Window 20 frames, with Window 30 frames achieving mean accuracy of 0.9031. The minimal difference (0.0000) between Window 20 and Window 30 suggests a performance plateau, where additional temporal context provides diminishing returns.

## 7.2 Why Window 30 Frames is Optimal

Based on the comprehensive analysis, Window 30 frames is recommended as the optimal window size for the following reasons:

1. **Highest F1 (Behavior) Score:** Window 20 frames achieves the highest F1 (Behavior) score of 0.8214, which is the most relevant metric for behavior classification. This score balances precision and recall for the behavior class, ensuring both accurate detection and minimal false positives.
2. **High Accuracy:** With a mean accuracy of 0.9031 across all videos, Window 30 frames provides excellent overall classification performance.
3. **Good Stability:** The standard deviation of 0.0882 for accuracy and 0.1870 for F1 (Behavior) indicates consistent performance across videos, which is important for reliable classification in production settings.
4. **Balanced F1 Scores:** Window 30 frames achieves good performance for both behavior class ( $F1 = 0.8087$ ) and not-behavior class ( $F1 = 0.9225$ ), indicating balanced classification across both classes.
5. **Computational Efficiency:** Compared to Window 30 frames, Window 30 frames provides similar performance with less computational overhead, making it more efficient for real-time or large-scale applications.

## 8 Discussion

### 8.1 Implications for Behavior Classification

The finding that larger window sizes (20-30 frames) outperform smaller windows (5-10 frames) has important implications for behavior classification:

- **Temporal Context Matters:** The *turn\_left* behavior requires sufficient temporal context to be accurately identified. This suggests that the behavior involves dynamic patterns that unfold over multiple frames, rather than being identifiable from single-frame or very short sequences.
- **Optimal Temporal Scale:** The performance plateau between Window 20 and Window 30 frames suggests that approximately 20 frames represents the optimal temporal scale for capturing *turn\_left* behavior. Beyond this, additional temporal context provides minimal benefit.
- **Window Size as Hyperparameter:** The significant performance differences across window sizes highlight the importance of window size as a critical hyperparameter that should be carefully tuned for each behavior class.

### 8.2 Limitations and Considerations

Several limitations should be considered when interpreting these results:

1. **Single Behavior Class:** This analysis focuses exclusively on *turn\_left* behavior. Results may not generalize to other behaviors, which may have different optimal window sizes.
2. **Feature Set:** The analysis uses a specific feature set. Different features may show different sensitivity to window size.

3. **Video Characteristics:** The optimal window size may vary depending on video characteristics such as frame rate, resolution, and behavioral context.
4. **Cross-Validation Structure:** The leave-one-animal-out cross-validation ensures independence but may not fully capture performance in scenarios with different animal distributions.

### 8.3 Future Directions

Several directions for future research and improvement are suggested by these findings:

- **Behavior-Specific Optimization:** Test window sizes for other behavior classes to determine if optimal window sizes are behavior-specific
- **Adaptive Window Sizing:** Develop methods that adapt window size based on video characteristics or behavioral context
- **Feature Engineering:** Investigate features that are less sensitive to window size while maintaining discriminative power
- **Ensemble Approaches:** Explore combining predictions from multiple window sizes to leverage the strengths of different temporal scales

## 9 Recommendations

### 9.1 Primary Recommendation

Based on the comprehensive analysis, we recommend using **Window 30 frames** for optimal performance in *turn\_left* behavior classification. This window size provides:

- Highest F1 (Behavior) score of 0.8214
- Excellent accuracy of 0.9031
- Good stability with standard deviation of 0.0882
- Balanced performance across both behavior and not-behavior classes

### 9.2 Alternative Considerations

If slightly higher accuracy is needed and stability is less critical, Window 30 frames (mean accuracy: 0.9031) may be considered, though the improvement is minimal (0.0000 difference) and comes with increased computational cost.

### 9.3 Data Quality Recommendations

- **Review Worst Performing Videos:** The top 10 worst-performing videos (Table 2) should be manually reviewed to identify potential data quality issues, annotation errors, or video quality problems.
- **Investigate Sensitive Videos:** Videos with high coefficient of variation (Table 3) should be examined to understand why they are sensitive to window size. This may reveal important behavioral patterns or inform feature engineering.
- **Validate Annotations:** Consider re-annotating a subset of worst-performing videos to ensure annotation quality and consistency.

## 9.4 Model Improvement Recommendations

- **Window Size Optimization:** For videos showing high sensitivity, test intermediate window sizes (e.g., 18, 22, 25 frames) to fine-tune performance.
- **Feature Analysis:** Examine feature importance rankings to identify which features are most critical for each window size, potentially informing feature engineering.
- **Video-Specific Tuning:** Consider developing video-specific or context-specific window size selection strategies for highly sensitive videos.

## 10 Conclusion

This comprehensive analysis of window size effects on *turn\_left* behavior classification reveals that Window 30 frames provides optimal performance, achieving a mean accuracy of 0.9031 and F1 (Behavior) score of 0.8087. The analysis demonstrates that larger window sizes (20-30 frames) significantly outperform smaller windows (5-10 frames), suggesting that *turn\_left* behavior requires sufficient temporal context for accurate classification.

The identification of worst-performing and window-sensitive videos provides actionable insights for data quality improvement and model optimization. Future work should explore behavior-specific window size optimization and adaptive window sizing strategies to further improve classification performance.

## Data Quality Notes

### Cross-Validation Structure

It is important to note that the cross-validation structure used in this analysis holds out **one animal at a time**, not one video at a time. Each video file contains multiple animals (identities), identified by numbers [0], [1], [2], etc. at the end of the video filename. Each (video\_name, identity) pair represents a separate test case, ensuring that model performance is evaluated independently for each animal. This structure is correctly handled throughout the analysis pipeline.

### Data Completeness

Some videos may be missing from certain window sizes, which may indicate data collection or processing issues. The validation report (available in `data/processed/validation_report.txt`) provides detailed information about data completeness and consistency.

## A Summary Statistics by Window Size

Table 4 provides complete summary statistics for all window sizes.

Table 4: Complete Summary Statistics

Window	Mean Acc.	SD Acc.	Mean F1 (Beh.)	SD F1 (Beh.)	Mean F1 (Not)	SD F1 (Not)
5	0.8234	0.0947	0.7249	0.1356	0.8475	0.1118
10	0.8717	0.1023	0.7926	0.1457	0.8818	0.1368
15	0.8927	0.0854	0.8073	0.1749	0.9094	0.0843
20	0.9026	0.0890	0.8214	0.1620	0.9170	0.0988
30	0.9031	0.0882	0.8087	0.1870	0.9225	0.0829

## B Data Files

All processed data files are located in `data/processed/`:

- `video_results.csv`: Complete performance data for each (video, identity, window) combination
- `summary_stats.csv`: Window-level summary statistics
- `feature_importance.csv`: Top features by importance for each window size
- `validation_report.txt`: Detailed validation results

## C Visualizations

Comprehensive visualizations including barbell plots, box-whisker plots, and lollipop plots are included in this report and are also available in the HTML report (`reports/window\_size\_analysis\report.html`).

These visualizations provide detailed graphical representations of the performance patterns described in this report.