Count What Repeats: Period-Adaptive Multi-Scale Consistency for Self-Supervised Repetitive Action Counting

Supplementary Material

7. Supplementary Experimental Results

In this section, we provide additional experimental results that were not included in the main paper due to space constraints. These results offer a more comprehensive evaluation of our proposed PAMS method, including a detailed comparison with state-of-the-art methods, a breakdown of counting accuracy, and further robustness analyses.

7.1. Off-By-N Accuracy Analysis

To provide a more fine-grained analysis of counting accuracy beyond the OBO metric, we report the Off-By-N (OBN) accuracy on the RepCount dataset in Table 3. This metric calculates the percentage of videos where the absolute difference between the predicted count and the ground-truth count is less than or equal to N ($|C_{pred} - C_{gt}| \leq N$). The results show that our model achieves over 90% accuracy for N=3, indicating high reliability.

Concurrently, the Cumulative Distribution Function (CDF) curve in Figure 7 confirms PAMS's superior accuracy. The curve shows that PAMS yields a higher fraction of videos within any given error tolerance compared to prior methods. For the critical tolerance of one absolute error (OBO metric), PAMS correctly counts 84.2% of videos on RepCount, demonstrating a substantial margin over prior work. Together, these plots illustrate the lower MAE and higher OBO accuracy achieved by our method.

Table 3. Off-By-N (OBN) accuracy on the RepCount dataset.

N	Accuracy on RepCount
0	0.5263
1	0.8421
2	0.8684
3	0.9013
4	0.9145
5	0.9145

7.2. Robustness to Variations in Video Speed

We conducted two experiments, not included in the main paper, to evaluate our model's robustness to tempo changes.

7.2.1. Internal Tempo Fluctuation

While global speed changes test robustness to uniform tempo, real-world scenarios often involve *non-stationary* tempo fluctuations, where an action may suddenly speed up or slow down. To simulate this more challenging condition, we conducted a stochastic perturbation experiment.

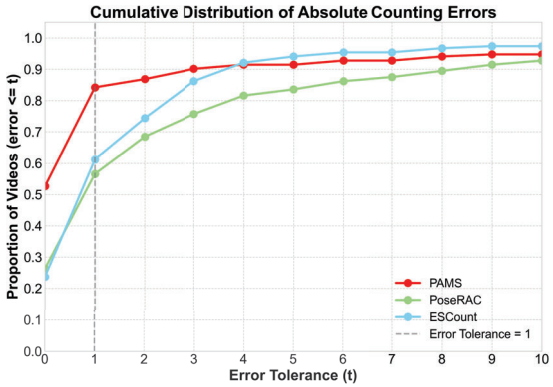


Figure 7. **Cumulative Distribution of Absolute Counting Errors.** The curve shows the fraction of videos (y-axis) whose absolute count error is less than or equal to a given value (x-axis). A curve shifted towards the top-left indicates better performance. PAMS achieves lower errors across the error tolerance range.

For each video in the RepCount test set (152 videos), we randomly selected a 5-7s segment. A random speed multiplier, drawn uniformly from $[0.5\times, 1.5\times]$, was applied exclusively to this segment, distorting the video's internal periodic structure. Due to the stochastic nature of both segment selection and speed ratio, we repeated this process 10 times for each video, resulting in 1,520 total trials.

The averaged results are presented in Table 4. Despite this challenging perturbation, the model's performance degrades gracefully. The average MAE increases only slightly from 0.1285 (clean) to 0.1336, and the OBO accuracy remains robust at 0.7447. This demonstrates a high degree of resilience to sudden, localized tempo drift.

Table 4. Average performance on RepCount against internal tempo fluctuations (a random 5-7s segment is speed up/slowed down 10 times per video).

MAE ↓	OBO ↑
0.1336	0.7447

Figure 8 and Table 5 provide a more detailed breakdown of this experiment by ground-truth count range. The analysis reveals that the performance degradation is primarily concentrated in videos with higher repetition counts (e.g.,

11+). In contrast, performance on low-count videos (0-10) remains almost entirely unaffected. This is an expected outcome, as a 5-7 second distortion introduces more significant periodic confusion in a high-density sequence. The model's stability in low-count ranges further confirms its robustness.

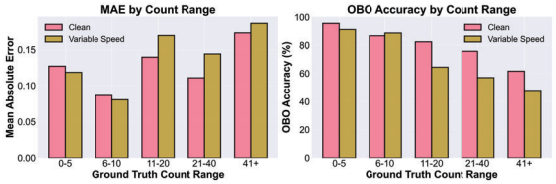


Figure 8. Performance breakdown by count range for the internal tempo fluctuation experiment. "Clean" refers to the original, unperturbed video performance. "Variable Speed" is the average result over 10 stochastic trials.

Table 5. Detailed MAE and OBO breakdown by count range for the internal tempo fluctuation experiment (averaged over 10 trials).

Count Range	Clean MAE ↓	Variable Speed MAE ↓	Clean OBO ↑	Variable Speed OBO ↑
0-5	0.1270	0.1184	0.9556	0.9111
6-10	0.0867	0.0807	0.8667	0.8867
11-20	0.1397	0.1694	0.8235	0.6441
21-40	0.1108	0.1442	0.7586	0.5690
41+	0.1730	0.1863	0.6154	0.4769

7.3. Performance with Limited Training Data

The main paper briefly mentions performance with 25% data (OBO of 0.663). Table 6 provides a detailed breakdown including MAE and results for 50% and 75% data subsets, showing strong performance even with significantly less data.

Table 6. Performance on RepCount when trained with varying amounts of training data.

Training Data %	MAE↓	OBO↑
25%	0.2458	0.6632
50%	0.1408	0.7612
75%	0.1355	0.8224
100%	0.1285	0.8421

7.4. Robustness Augmentation Parameters

For clarity and reproducibility, we provide the detailed augmentation parameters employed in the robustness experiments summarized in Table 3 of the main paper. The augmentation pipeline incorporates two distinct transformation types, each with carefully calibrated parameters to evaluate model resilience under challenging conditions.

The **occlusion augmentation** introduces partial observation scenarios by randomly masking portions of the input sequences. This transformation employs a time ratio of 0.2, controlling the proportion of temporal segments to be occluded, alongside a joint ratio of 0.3, governing the fraction of body joints to be masked at each affected timeframe.

Complementarily, the **affine transformation** simulates viewpoint variations and spatial perturbations through geometric manipulations. This includes random rotations within a $\pm 15^\circ$ range, scaling operations with a variation factor of 0.15 around the original scale, and spatial translations constrained to 0.1 of the coordinate range along each axis.

These parameter values were empirically determined to create meaningful distribution shifts while preserving the semantic integrity of human motion patterns, thereby enabling a comprehensive assessment of model robustness against realistic perturbations.

7.5. Robustness to Occlusion and Affine Transformations

In addition to the augmentation parameters specified, we report the quantitative results for these robustness tests on the RepCount dataset in Table 7. These results, referenced in the main paper, confirm the model's stability against common visual corruptions. 'Occlusion' involves randomly masking 30% of joints for 20% of the time, while 'Affine' involves applying random rotation ($\pm 15^\circ$), scaling ($\pm 15\%$), and translation ($\pm 10\%$).

Table 7. Performance on RepCount under occlusion and affine transformations.

Transformation	MAE ↓	OBO ↑
Occlusion	0.1366	0.7895
Affine (View)	0.1220	0.8289

7.6. Robustness to Frame Dropping and Duplication

We conducted an additional analysis to evaluate model robustness against common video stream corruptions. We simulated two scenarios: random frame dropping (removal) and random frame duplication (repetition). Both augmentations were applied to the RepCount test set at varying rates of 5%, 10%, and 20%.

The results are presented in Figure 9 (frame dropping) and Figure 10 (frame duplication). Both plots demonstrate that the model maintains strong and stable performance. The MAE and OBO metrics exhibit only minor fluctuations, even when 20% of the video frames are corrupted. This analysis confirms the model's high resilience to common stream quality issues, such as those caused by network jitter or packet loss.

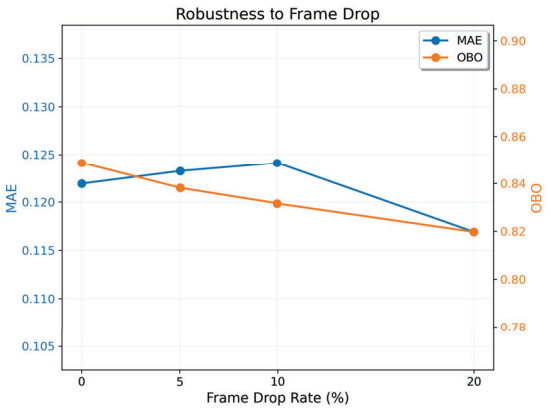


Figure 9. Performance on RepCount under random frame dropping. MAE (left y-axis) and OBO (right y-axis) remain stable as the drop rate increases, showing robustness to missing frames.

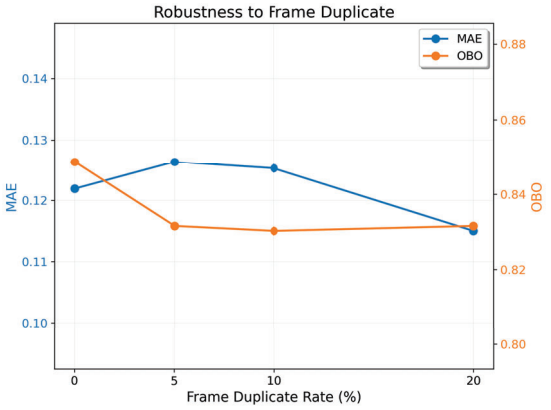


Figure 10. Performance on RepCount under random frame duplication. The model shows high resilience, with minimal degradation in MAE or OBO, indicating it is not sensitive to redundant frames.

7.7. Leave-One-Out (LOO) Cross-Category Generalization

To assess the generalization of our learned periodic representation, we conducted a leave-one-out (LOO) cross-category analysis. We trained 8 separate models, each time excluding one specific action category from the RepCount training set. We then evaluated each model's zero-shot performance on the single category that was held out during its training.

This experiment tests the model's ability to count a completely unseen action, relying only on its general understanding of periodic motion. The results are presented in Table 8. The model achieves strong performance on most unseen categories, such as jump_jack (0.0354 MAE) and pommelhorse (0.0323 MAE). Performance was most challenged by bench_pressing (0.2509 MAE), which has subtle motion cues.

Overall, the strong results confirm that our PAMS TCC objective learns a generalizable representation of periodicity, rather than simply memorizing the patterns of specific, seen action types.

Table 8. Leave-One-Out (LOO) zero-shot performance on RepCount. Each row shows the performance on an action category that was excluded from the training data.

Held-Out Action	MAE ↓	OBO ↑
front_raise	0.1121	0.9286
pull_up	0.1172	0.7500
squat	0.0971	0.6923
bench_pressing	0.2509	0.6250
jump_jack	0.0354	0.9130
situp	0.1417	0.9000
push_up	0.1886	0.8236
pommelhorse	0.0323	0.9333

7.8. Prediction Agreement Analysis (Bland-Altman)

To further evaluate the model's performance beyond error metrics, we conducted a Bland-Altman analysis to assess the agreement between the model's predicted counts and the ground-truth counts. This method plots the difference between the two measurements against their mean, providing insight into systematic bias and the range of agreement.

As shown in Figure 11, the analysis reveals a mean bias of -1.3576 (95% CI: [-2.5169, -0.3841]). This indicates a small but statistically significant systematic tendency for the model to slightly underestimate the true count. The 95% limits of agreement (LoA) are [-14.4771, 11.7619], which define the range where 95% of the differences are expected to fall. While this range appears wide, the plot shows that the vast majority of predictions are tightly clustered around the zero-difference line, especially for videos with lower repetition counts (e.g., <40), which form the bulk of the dataset. The outliers primarily occur in high-count videos, which is consistent with the OBN analysis. Overall, the plot confirms strong agreement for most of the data and identifies a minor systematic underestimation.

7.9. Hyperparameter Sensitivity Analysis (Training)

To complement the inference analysis, we also evaluated the model's sensitivity to key hyperparameters during the self-supervised training phase. This analysis validates that the parameters chosen for our final model (which achieved MAE 0.129 and OBO 0.842 on RepCount) are located

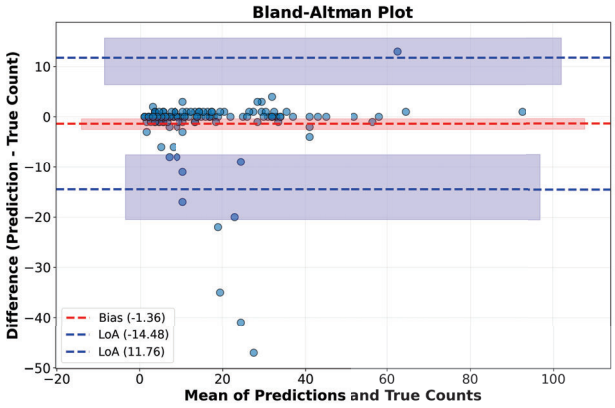


Figure 11. Bland-Altman plot for agreement between predicted and ground-truth counts on the RepCount dataset (N=152). The red dashed line shows the mean bias (-1.36), with its 95% CI in the shaded red area. The blue dashed lines represent the 95% limits of agreement (LoA) [-14.48, 11.76], with their CIs in the shaded blue areas.

within a stable and robust performance region. We evaluated four critical parameters: TCC Window Size, TCC Stride, TCC Temperature, and the Cross-Class Weight.

The results are presented in Figure 12.

- **TCC Window Size (Figure 12a):** The plot shows stable MAE and OBO performance across various window sizes. Our chosen model's parameters fall within this robust range, and deviations (e.g., to 20 or 30) result in only minor, acceptable fluctuations.
- **TCC Stride (Figure 12b):** Performance remains strong for strides between 2 and 6. This validates our parameter choice as being in a region of high OBO and low MAE, demonstrating robustness to this sampling interval.
- **TCC Temperature (Figure 12c):** The InfoNCE temperature τ (set to 0.1 in our main paper) is shown to be a robust choice. While tuning τ causes fluctuations (e.g., ± 0.01 MAE), the performance across this range remains high, confirming that the model is not overly sensitive to this value.
- **Cross-Class Weight (Figure 12d):** This analysis shows that performance is robust to the weighting of hard negative samples. Our chosen parameter is validated as effective, with other weights yielding similar, stable performance within a small range.

Overall, these experiments confirm that our training objective is not overly sensitive to individual hyperparameter settings. The parameters used for our final model are shown to be in a robust, high-performance region, with variations causing only small and acceptable fluctuations.

Finally, we analyzed the correlation between these training hyperparameters and the final performance metrics, as shown in Figure 13. The matrix reveals several insights. For instance, TCC Temperature shows a moderate positive correlation with OBO (0.331) and a negative correlation with MAE (-0.212). Similarly, TCC Stride has a negative correlation with MAE (-0.318). As expected, MAE and OBO themselves show a strong negative correlation (-0.487). This analysis quantifies the relative impact of each parameter, reinforcing that they are meaningful levers for tuning performance.

7.10. Hyperparameter Sensitivity Analysis (Inference)

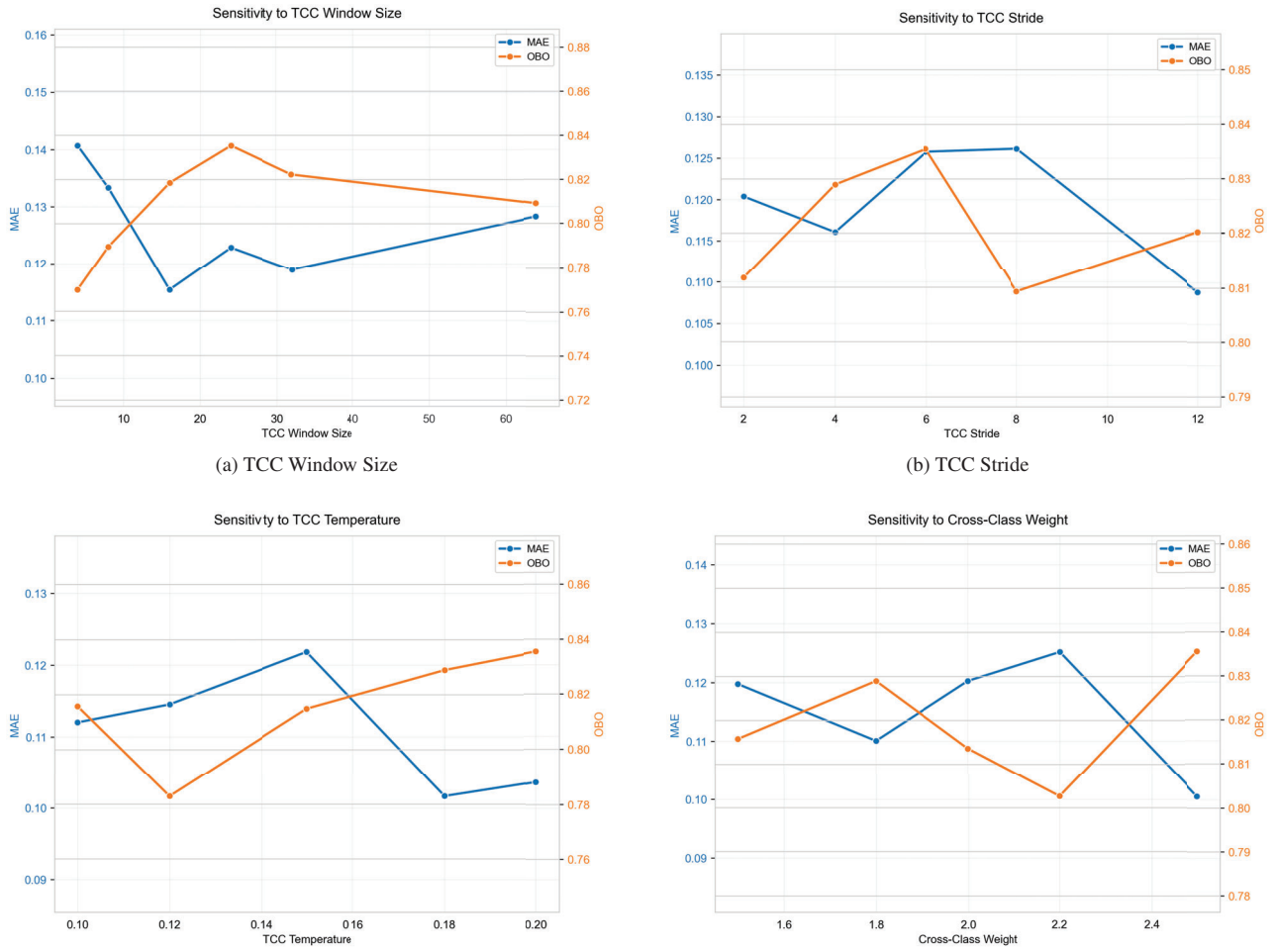
We analyzed the model's sensitivity to key hyperparameters in the inference algorithm to assess its robustness to parameter choices. We evaluated five crucial parameters of our multi-expert counting algorithm, which is based on peak detection: the 'First Sigma Multiplier' and 'First Distance Multiplier' (which control the smoothing and peak spacing for the 'fast' expert), the 'Height Factor' and 'Prominence Factor' (which control the peak detection thresholds), and the 'Long Window Weight' (affecting the analysis window or feature contribution during inference).

The results are shown in Figures 14 through 18. Each plot demonstrates that both MAE and OBO metrics remain relatively stable across a reasonable range of values for the corresponding parameter. Specifically, Figure 14 shows the sensitivity to the First Sigma Multiplier, Figure 15 examines the First Distance Multiplier, Figure 16 analyzes the Height Factor, Figure 17 evaluates the Prominence Factor, and Figure 18 investigates the Long Window Weight. These results indicate that our multi-expert consensus algorithm is not overly sensitive to fine-tuning and is robust to parameter variations, which is a desirable property for real-world deployment.

Furthermore, to ensure that the model is not sensitive to the *interaction* between hyperparameters, we conducted a series of joint sensitivity analyses. The following figures evaluate coupled parameter effects across different expert configurations and global detection parameters.

Figures 19, 20, and 21 show the joint sensitivity for the 'Sigma Multiplier' (smoothing) and 'Distance Multiplier' (peak spacing) for each of the three experts (fast, medium, slow), respectively. Each figure presents dual heatmaps showing MAE (left) and OBO (right) metrics across the tested parameter ranges. The heatmaps remain stable across all three experts, indicating that the performance of our multi-expert system is not dependent on a specific, finely-tuned pairing of these expert-specific parameters.

The analysis of global peak detection parameter interactions is presented in Figures 22, 23, and 24, which evaluate the interactions between 'Height Factor', 'Prominence Factor', and 'Long Window Weight'. Figure 22 examines the Height Factor versus Prominence Factor interaction, Figure 23 analyzes Prominence Factor versus Long Window



(a) TCC Window Size

(b) TCC Stride

(c) TCC Temperature

(d) Cross-Class Weight

Figure 12. Sensitivity analysis for key **training** hyperparameters on the RepCount dataset. We evaluate (a) TCC Window Size, (b) TCC Stride, (c) TCC Temperature, and (d) Cross-Class Weight. The plots show stable performance fluctuations around our final model's parameters.

Weight, and Figure 24 investigates Height Factor versus Long Window Weight. These results demonstrate a high degree of robustness. While extreme combinations can lead to minor performance drops (e.g., very high 'Prominence Factor' combined with high 'Height Factor'), there is a large, stable region of low MAE and high OBO across all parameter pairs.

This comprehensive sensitivity analysis reinforces the conclusion that our multi-expert inference algorithm is robust to both individual parameter choices and their interactions, a desirable trait for practical deployment.

7.11. End-to-End Performance and Real-Time Analysis

To evaluate the practical deployment viability of our method, we analyzed the end-to-end performance of the full pipeline. This pipeline includes both the upstream pose extraction stage and our proposed PAMS inference stage.

Our analysis shows that the system achieves a mean end-to-end processing speed of **24.81 FPS**, with a median of 26.19 FPS. The full distribution is shown in Figure 25. This average speed is nearly identical to the 25 FPS real-time target, which is a common frame rate for video capture in many real-world scenarios. This result demonstrates that our model is capable of providing stable, real-time counting for typical life and fitness applications.

We further analyzed the computational cost of each component, as shown in Figure 26. The average time to process a single frame is 42.03 ms. This total time is dominated by the pose extraction step, which requires 25.81 ms (61.4% of the total time). In contrast, our PAMS inference algorithm is highly efficient, requiring only **16.22 ms** (38.6% of the total

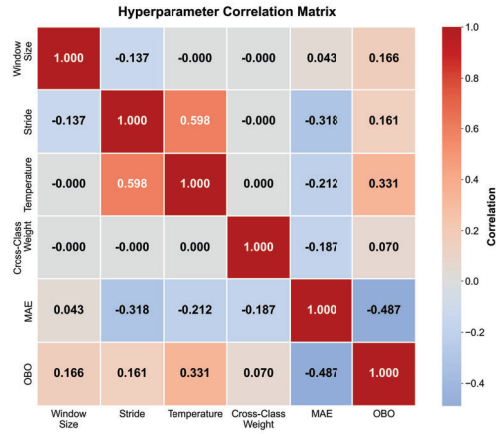


Figure 13. Correlation matrix for key **training** hyperparameters against final performance metrics (MAE and OBO) on RepCount. This analysis helps quantify the relative impact of each parameter on the model's accuracy.

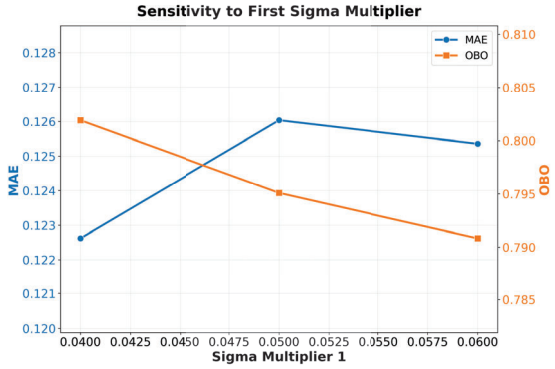


Figure 14. Sensitivity analysis for the First Sigma Multiplier parameter of the inference algorithm on the RepCount dataset. The plot shows stable MAE and OBO performance across a range of values, demonstrating robustness to this smoothing parameter for the fast expert.

time). This clearly indicates that the primary performance bottleneck is pose extraction, not our counting algorithm.

The system's real-time capability (where processing FPS $\geq$ video FPS) is strongest for videos at common frame rates. For instance, the system can process 100% of videos recorded at <20 FPS and 77.1% of videos recorded between 20-25 FPS in real-time. This confirms its suitability for widespread use. Furthermore, the model is lightweight, with a mean peak VRAM usage of only 100.91 MB, making it easy to deploy on a wide range of devices.

7.12. Qualitative Analysis of Challenging Scenarios

To further demonstrate the robustness of our method, we provide qualitative results on challenging real-world scenarios. These examples test the model's resilience to significant viewpoint changes and interruptions in motion.

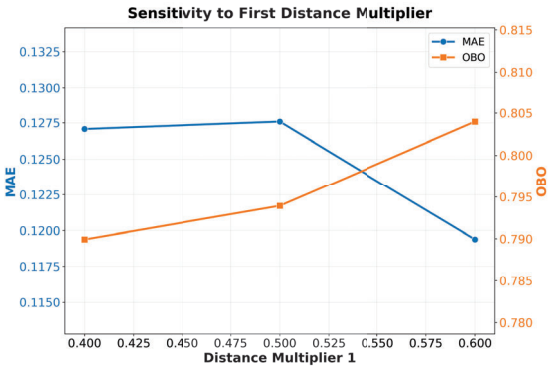


Figure 15. Sensitivity analysis for the First Distance Multiplier parameter of the inference algorithm on the RepCount dataset. Performance remains consistent across different peak spacing values for the fast expert.

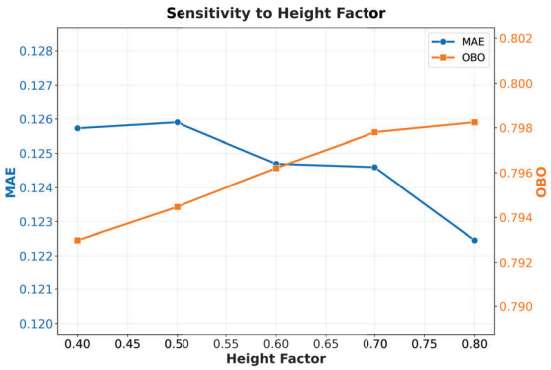


Figure 16. Sensitivity analysis for the Height Factor parameter of the inference algorithm on the RepCount dataset. The model demonstrates robustness to variations in peak detection height thresholds.

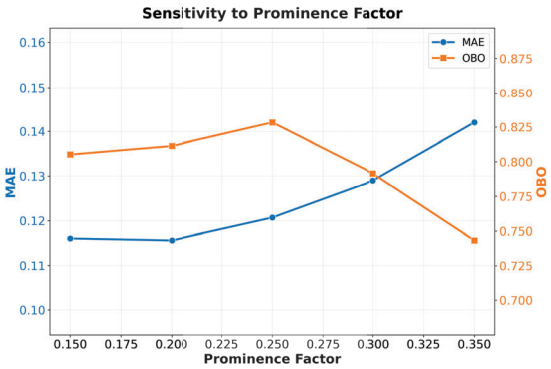


Figure 17. Sensitivity analysis for the Prominence Factor parameter of the inference algorithm on the RepCount dataset. Performance is stable across different prominence threshold settings.

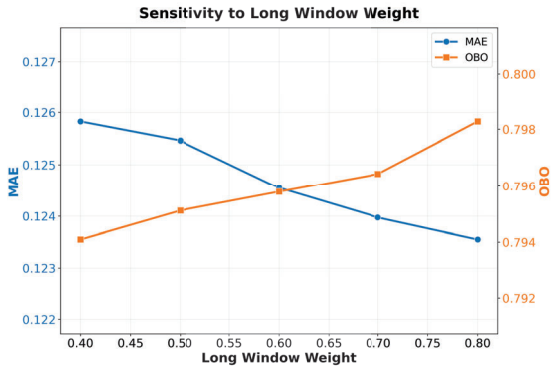


Figure 18. Sensitivity analysis for the Long Window Weight parameter of the inference algorithm on the RepCount dataset. The model shows consistent performance across different window weighting configurations.

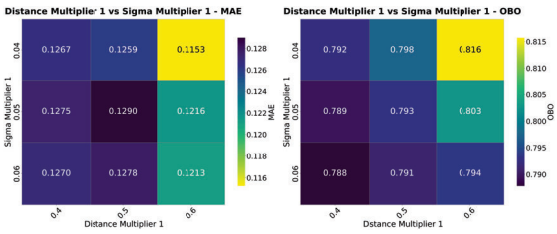


Figure 19. Joint sensitivity analysis for Expert 1 (Fast) on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different combinations of Sigma Multiplier and Distance Multiplier. Performance remains stable across the tested parameter ranges.

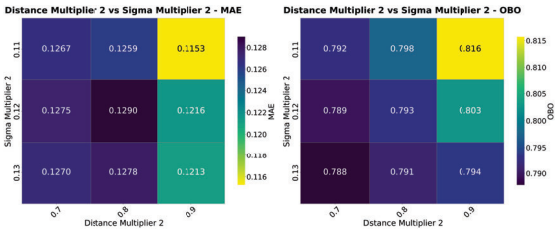


Figure 20. Joint sensitivity analysis for Expert 2 (Medium) on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different combinations of Sigma Multiplier and Distance Multiplier. Consistent with Expert 1, the performance exhibits high stability.

7.12.1. Successful counting example

Figure 27 demonstrates the model's performance on a long sequence with substantial viewpoint variation. The subject performs 58 repetitions while the camera angle changes. Our model's feature representation remains stable. The multi-expert inference algorithm successfully tracks the periodicity, achieving a perfect count of 58. This highlights robustness to both long sequences and view changes.

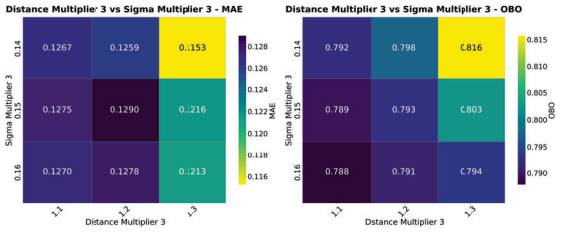


Figure 21. Joint sensitivity analysis for Expert 3 (Slow) on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different combinations of Sigma Multiplier and Distance Multiplier. The robust performance pattern is maintained across all three experts.

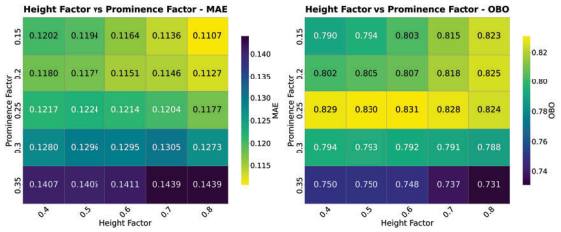


Figure 22. Joint sensitivity analysis for Height Factor versus Prominence Factor on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different parameter combinations. The model demonstrates robustness with a large stable region of low MAE and high OBO.

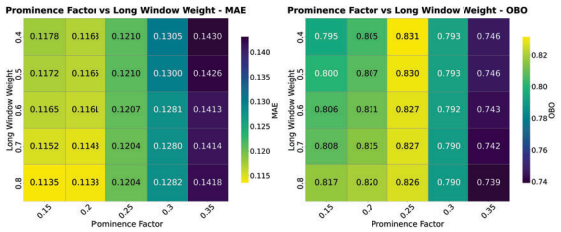


Figure 23. Joint sensitivity analysis for Prominence Factor versus Long Window Weight on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different parameter combinations. Performance remains stable across most of the parameter space.

Figure 28 illustrates robustness to aperiodic interruptions. The subject performs 17 repetitions but pauses in the middle of the video (approximately frames 800-1000). During this pause, the motion curve flattens, and no periodic signal is present. Our adaptive peak detection algorithm correctly identifies this non-periodic segment. It does not generate false positive counts during the pause. The model accurately predicts the ground truth of 17, demonstrating its ability to handle interruptions.

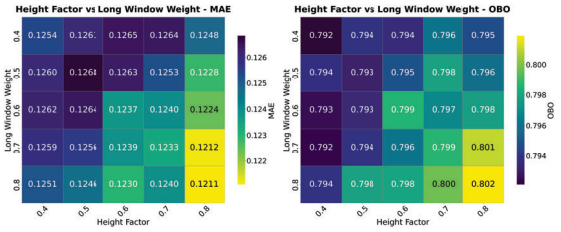


Figure 24. Joint sensitivity analysis for Height Factor versus Long Window Weight on the RepCount dataset. The dual heatmaps show MAE (left) and OBO (right) across different parameter combinations. The analysis confirms robustness with consistent performance across the tested ranges.

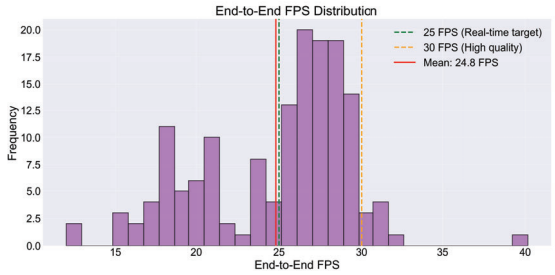


Figure 25. End-to-end FPS distribution over 152 test videos. The mean performance of 24.8 FPS (red line) is very close to the 25 FPS real-time target (green dashed line), indicating strong performance for common scenarios.

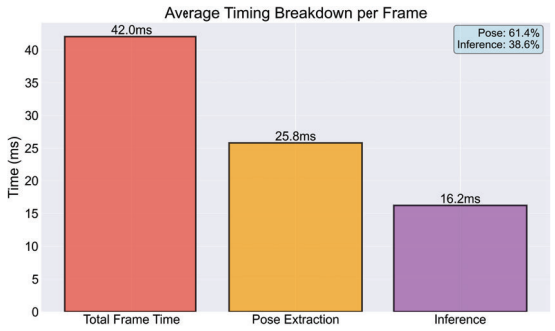


Figure 26. Average timing breakdown per frame. Pose extraction is the main computational bottleneck, accounting for 61.4% of the processing time. Our inference model is efficient, using only 38.6% of the time.

7.12.2. Failure count example

To thoroughly discuss the limitations and applicable scenarios of our model, we also provide examples where the model fails and analyze the reasons for inaccurate counting. Figure 29 shows the model's performance on the sequence with the largest counting error. In this video, the subject is far from the camera, resulting in a small observed motion amplitude, which makes it difficult to generate a motion curve with clear peaks and troughs. After algorithmic smoothing, the curve representing the repetitive action is flattened, making it hard for the algorithm to distinguish between action states and leading to an incorrect count. This indicates that our model still has limitations when dealing with actions of very small amplitude. However, since such scenarios rarely occur in real-world production and daily life—and the issue can be mitigated by zooming in or focusing the camera on the subject—our model remains valuable and promising for practical applications.

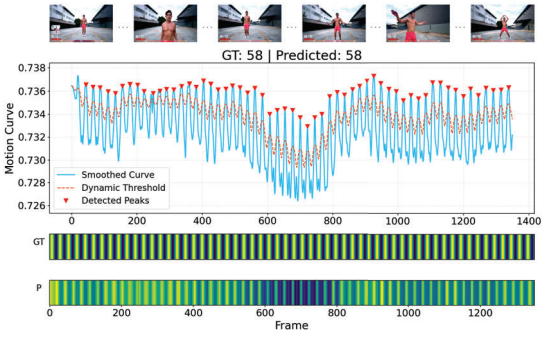


Figure 27. Robustness to viewpoint changes and long sequences. The model accurately processes a long video (GT: 58, Predicted: 58) despite significant shifts in camera angle. The motion curve and dynamic threshold adapt correctly, leading to a precise prediction of 58.

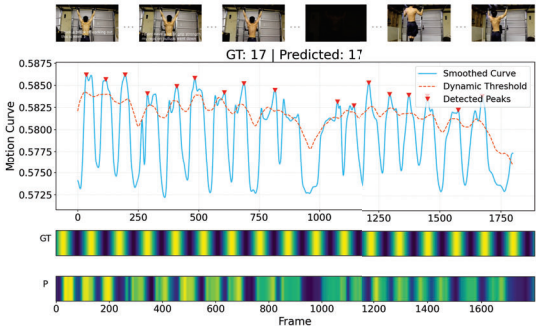


Figure 28. Robustness to aperiodic interruptions. This example shows an action sequence with a distinct pause (approx. frames 800-1000). Our method successfully identifies this non-periodic segment and prevents false positives. The model correctly counts only the true repetitions (GT: 17, Predicted: 17).

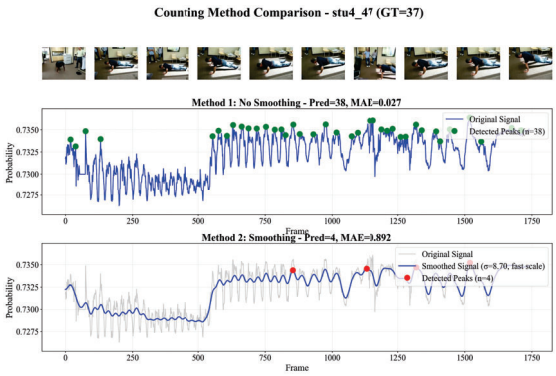


Figure 29. This example illustrates the model's incorrect counting when faced with actions that have very small amplitude and low frequency. Due to smoothing, the original peaks in the motion curve are lost, causing the counting mechanism to fail. When smoothing is disabled, accurate counting can be restored.