A Machine Learning Framework for Automated Computational Ethology Using Markerless Pose Estimation

Prateek Verma
EECS, Univ. of Arkansas
Fayetteville, AR 72701
prateek@uark.edu

Minh-Hao Van
EECS, Univ. of Arkansas
Fayetteville, AR 72701
haovan@uark.edu

Christopher McEnaney
BISC, Univ. of Arkansas
Fayetteville, AR 72701
cm109@uark.edu

Ashtyn Le
Fayetteville High School
Fayetteville, AR 72701
ashtyn.nalin@gmail.com

Amy R. Poe
BISC, Univ. of Arkansas
Fayetteville, AR 72701
amypoe@uark.edu

Xintao Wu
EECS, Univ. of Arkansas
Fayetteville, AR 72701
xintaowu@uark.edu

Abstract

Quantitative behavioral analysis is fundamental to ethological research, yet automated approaches remain limited by the gap between pose estimation and meaningful behavioral classification. Most existing methods focus on either pose detection or behavior recognition in isolation, lacking integrated frameworks for comprehensive behavioral analysis. We present an end-to-end framework that bridges markerless pose estimation with machine learning classification for automated behavioral analysis. Our framework integrates SLEAP pose estimation, systematic feature engineering, multiple machine learning algorithms, and robust validation strategies into a unified pipeline. We demonstrate the framework on Drosophila larvae videos, automatically classifying three behavioral states (feeding, sleeping, crawling) from pose trajectories. We evaluate five machine learning models across three validation strategies and engineer twelve position-invariant features from four anatomical landmarks. The framework provides computational ethology researchers with practical tools for pose-based behavioral classification, comprehensive model evaluation, and deployment guidance for real-world applications.

1 Introduction

Quantitative analysis of animal behavior represents a cornerstone of ethological research, providing insights into evolutionary adaptations, ecological relationships, and neurobiological mechanisms underlying complex behaviors. Traditional behavioral analysis methods rely heavily on manual observation and coding processes that are time-intensive, subject to observer bias, and limited in their capacity to capture subtle behavioral nuances or analyze large-scale datasets. The emergence of computational ethology has introduced automated approaches that promise to revolutionize behavioral research through objective, high-throughput analysis. Accurate behavioral prediction is particularly valuable for research on learning, memory formation, sensory perception, and abnormality detection. This is especially critical during development, when organisms exhibit rapid transitions between conflicting behavioral states (feeding for growth versus sleep for neural development) that later consolidate into mature circadian patterns [1, 2, 3]. However, current automated approaches face significant limitations, especially for developing organisms. Existing methods for larval behavior analysis, often based on heuristic image comparisons, are slow and limited in scalability. Most

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

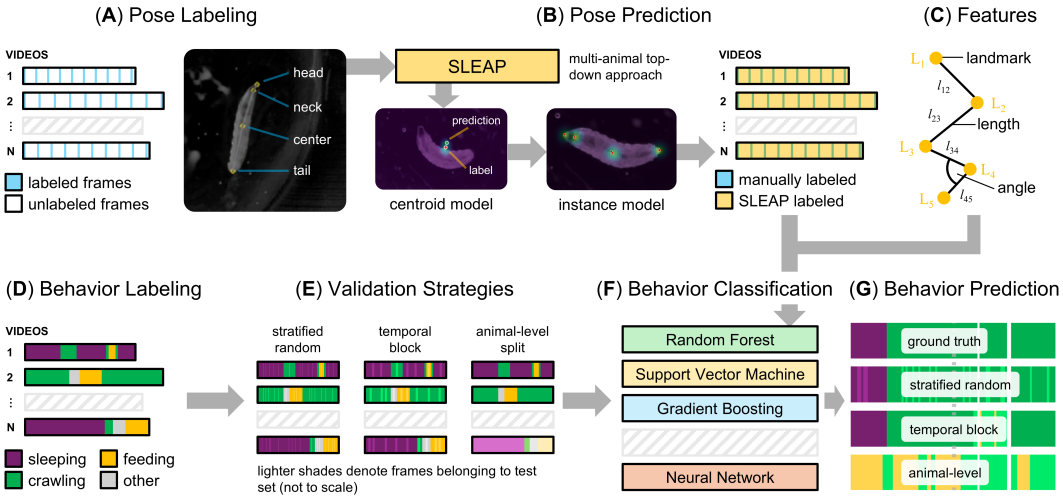

Figure 1: A machine learning powered framework for automated animal behavior classification using markerless pose estimation. (A) A few frames per video are labeled manually with animal pose (body parts of interest). (B) A markerless pose estimation tool (like SLEAP) is used to predict the poses for all remaining frames. (C) Pose features are processed and normalized. (D) Videos are labeled with the starting and ending frame number for each behavior. (E) A variety of validation strategies for splitting the frames into train and test sets are adopted. (F) Various ML behavior classifiers are trained. (G) A visualization for predicted behaviors is generated for each animal.

computational approaches focus on mature organisms with well-defined anatomical features, creating gaps for developing organisms with less distinct morphological landmarks (body parts).

Recent advances in computer vision and machine learning have enabled sophisticated tools for automated behavioral analysis, particularly markerless pose estimation technologies [4, 5]. However, a critical gap exists in translating pose trajectories into meaningful behavioral classifications. Most studies focus either on pose estimation accuracy or behavioral classification in isolation, rather than developing integrated end-to-end frameworks. For soft-bodied organisms like Drosophila larvae, existing approaches are limited in behavioral scope, typically focusing on locomotion while lacking comprehensive frameworks for simultaneous analysis of multiple behavioral states including sleeping, feeding, and social behaviors. Drosophila larvae represent an ideal model system due to their simplified neural architecture and established role in studying neurodevelopmental processes [2]. This study addresses these limitations by presenting a comprehensive framework for automated behavioral classification that integrates SLEAP pose estimation with machine learning approaches. We bridge the gap between pose estimation and behavioral classification through an end-to-end pipeline including pose labeling, feature engineering, multi-model evaluation, and robust validation strategies. Our framework is demonstrated on Drosophila larvae, enabling simultaneous classification of feeding, sleeping, and crawling behaviors while providing practical guidance for researchers. Through systematic evaluation of validation strategies, we establish practices for temporal dependency handling and cross-individual generalization assessment essential for computational ethology applications.

Related work: Deep learning-based pose estimation has revolutionized animal behavior analysis through high-throughput, objective quantification. DeepLabCut [4] pioneered transfer learning for markerless pose estimation in laboratory animals, while SLEAP [5] extended these capabilities to multi-animal scenarios. Such systems employ convolutional neural networks based on U-Net [6], ResNet [7], or EfficientNet [8] architectures to predict anatomical landmark coordinates. Contemporary systems demonstrate varying approaches to automated phenotyping. FlyVISTA [9] utilizes closed-loop video imaging for adult Drosophila, providing higher spatiotemporal resolution than traditional DAM (Drosophila activity monitoring) systems [10] and enabling detection of microbehaviors like antennal movements during sleep. Similar advances in mice [11, 12] have established automated phenotyping capabilities for mature organisms with well-defined anatomical features. For Drosophila larvae, computational approaches remain limited. LarvaTagger [13], TrackMate [14], and IMBA [15] provide body position tracking but require complex scripting and focus primarily

on locomotion behaviors (crawling, bending, rolling). These systems lack integrated frameworks for comprehensive behavioral analysis encompassing sleeping [16, 2], feeding [17], and memory formation [18]. The soft-bodied morphology and limited anatomical landmarks of larvae present unique computational challenges that existing approaches have not fully addressed.

2 Framework

Our methodology comprises a comprehensive framework for automated behavioral classification using SLEAP pose estimation followed by machine learning classification. The pipeline consists of four main phases illustrated in Figure 1: (1) dual annotation of pose landmarks and behavioral states by domain experts, (2) training SLEAP models for automated pose prediction across all video frames, (3) engineering position-invariant features from pose coordinates, and (4) evaluating multiple machine learning algorithms with different validation strategies. We manually annotated 10 videos containing 13 Drosophila larvae with four anatomical landmarks (head, neck, center, tail) and three behavioral states (sleeping, feeding, crawling). A two-stage SLEAP pipeline automatically predicted pose coordinates across all frames. We engineered 12 features: 6 center-normalized coordinates, 4 inter-landmark distances, 1 body curvature measure, and 1 velocity-based activity metric. Five machine learning classifiers (Random Forest, SVM, Gradient Boosting, K-Nearest Neighbors, Neural Network) were evaluated using three validation strategies: stratified random split, temporal block-based split, and animal-level split. Detailed methodology is provided in Appendix A.

3 Results and discussion

The processed dataset comprises 16,552 frames from 10 videos featuring 13 Drosophila larvae, with behavioral distribution of: feeding (55.4%), sleeping (23.1%), and crawling (21.6%). Individual animals exhibited distinct behavioral profiles ranging from feeding-dominant to sleeping-dominant and crawling-dominant phenotypes, providing diverse training examples for machine learning classification. Dataset composition and animal profiles are provided in Appendix Tables 4 and 5.

Pose estimation performance: The centroid detection model finished training at 29 epochs, achieving a final training loss of 1.97×10^{-6} and validation loss of 1.22×10^{-5}. The centered instance model finished training at 18 epochs with a final training loss of 1.65×10^{-5} and validation loss of 2.49×10^{-4}. Both models demonstrated robust convergence during training with automatic learning rate reduction and early stopping triggered by plateau detection. They exhibited consistent decrease in validation loss throughout training, indicating effective learning without overfitting. Trained models achieved a mean pixel error of 3.05×10^{-5} and 2.12, and a mean percentage of correct keypoints (PCK) score of 100 and 82.17 for the centroid and centered-instance models respectively,

Table 1: Model performance results

		Stratified Random		**Temporal Blocks**		**Animal-Level**	
Feature Set	**Model**	Acc.	F1-score	Acc.	F1-score	Acc.	F1-score
Relative Coordinates (6 total)	RF	0.979	0.979	0.910	0.909	0.170	0.204
	SVM	0.811	0.806	0.827	0.821	0.187	0.262
	GB	0.862	0.857	0.845	0.839	0.196	0.257
	KNN	0.966	0.966	0.912	0.910	0.223	0.277
	NN	0.855	0.856	0.832	0.831	0.122	0.157
Relative Coordinates + Distances (10 total)	RF	0.977	0.977	0.913	0.912	0.131	0.152
	SVM	0.843	0.841	0.839	0.837	0.214	0.283
	GB	0.887	0.884	0.873	0.870	0.171	0.226
	KNN	0.945	0.945	0.897	0.895	0.244	0.308
	NN	0.848	0.851	0.868	0.867	0.134	0.162
Relative Coordinates + Distances + Activity + Curvature (12 total)	RF	**0.982**	**0.982**	**0.936**	**0.935**	0.224	0.269
	SVM	0.873	0.873	0.882	0.880	0.729	0.691
	GB	0.893	0.892	0.890	0.888	0.258	0.305
	KNN	0.952	0.952	0.909	0.906	0.696	0.697
	NN	0.889	0.890	0.890	0.890	**0.733**	**0.700**

demonstrating that the models found the identification of the position of the larva in the frame an easier task than the identification of their pose landmarks. This is expected for the larvae since they possess less distinct body parts when compared to adult Drosophila or other animals.

Behavioral classification performance: Model performance was evaluated across three validation strategies and feature sets to comprehensively assess generalization capabilities and feature importance. Results demonstrate clear performance differences between validation strategies: stratified random split achieves optimistic performance (98.2% accuracy for Random Forest with all features) due to temporal correlations, temporal block split provides more realistic estimates (93.6% accuracy with all features), and animal-level split reveals the most challenging cross-individual generalization scenario (73.3% accuracy for Neural Network with all features). The temporal block split successfully addresses temporal dependency by creating 15-frame chunks with balanced representation (831 training blocks and 195 test blocks). For the animal-level split, two animals with distinct behavioral profiles were chosen as test subjects: one exhibiting pure feeding behavior (1647 frames, 100% feeding) and another showing mixed behaviors (1265 frames: 53% feeding, 28% sleeping, 19% crawling), constituting 17.6% of the total dataset. This combination provides a realistic test for cross-individual generalization with feeding bias but behavioral variation. Feature engineering demonstrates consistent value across all validation strategies: coordinate normalization provides robust baseline performance, while adding distance features generally improves classification accuracy. Random Forest achieved the highest performance across feature sets for stratified random and temporal block splits, while Neural Network performed best for animal-level splits. Appendix Figure 2 shows the timeline for predictions, and Appendix Figure 3 the confusion matrices, for the best performing model for each validation strategy.

Framework validation and generalizability: The SLEAP pose estimation achieved 100% PCK for centroid detection and 82.17% PCK for instance models, validating the two-stage pipeline for challenging soft-bodied organisms. The systematic performance degradation from stratified random (98.2%) through temporal blocks (93.6%) to animal-level split (73.3%) successfully exposes different generalization challenges, confirming the framework's capacity to provide realistic performance assessment across deployment scenarios. The validation hierarchy reveals critical insights: temporal block splitting addresses temporal dependency while maintaining behavioral context, and animal-level evaluation exposes the substantial challenge of cross-individual generalization. Feature engineering proved robust across all validation strategies, with the 12-feature approach showing consistent patterns that validate the design for position invariance and behavioral transferability. Importance of proper model selection emerged clearly, demonstrating that one model could be optimal for within-video scenarios and another for cross-individual applications. The modular design enables adaptation across organisms and contexts while maintaining methodological rigor. This framework provides ethology researchers with bespoke evidence-based tools for integrating pose and behavior analysis.

4 Conclusions, limitations and future direction

Our framework integrates SLEAP-based pose estimation, feature engineering, multi-model evaluation, and robust validation strategies into a cohesive pipeline prioritizing methodological rigor and practical applicability. The pose estimation component demonstrates effective adaptation to challenging soft-bodied organisms through a two-stage pipeline, while the behavioral classification component establishes methodology for translating pose trajectories into behavioral insights. Evaluation across validation strategies reveals critical insights: stratified random splits provide optimistic baselines but suffer from temporal correlations, temporal block splits offer realistic performance estimates by addressing temporal dependencies, while animal-level splits expose the substantial challenge of cross-individual generalization. The 12-feature engineering approach establishes principles for extracting behaviorally meaningful descriptors that maintain position invariance and transferability. Experiments with a variety of models guide and emphasize the need for selecting the correct model for a given scenario. Current approach is limited to pose-based features; incorporating raw image patches or optical flow could capture behavioral nuances not represented in skeletal coordinates. Dataset size and class imbalance limit generalization capabilities. The framework's dependence on manual annotation creates scaling bottlenecks. Future research should integrate temporal sequence models for time-dependent behavioral patterns, develop semi-automated learning approaches to reduce annotation burden, and explore multi-modal integration.

Acknowledgment

This work was supported in part by the National Science Foundation under awards 1946391, the National Institute of General Medical Sciences of National Institutes of Health under award P20GM139768, and the Arkansas Integrative Metabolic Research Center at University of Arkansas; by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission; and by the American Academy of Sleep Medicine Foundation Bridge to Success Grant to A.R.P.

References

[1] Wenyu Huang, Kathryn Moynihan Ramsey, Biliana Marcheva, Joseph Bass, et al. Circadian rhythms, sleep, and metabolism. *The Journal of clinical investigation*, 121(6):2133–2141, 2011.

[2] Amy R Poe, Lucy Zhu, Milan Szuperak, Patrick D McClanahan, Ron C Anafi, Benjamin Scholl, Andreas S Thum, Daniel J Cavanaugh, and Matthew S Kayser. Developmental emergence of sleep rhythms enables long-term memory in drosophila. *Science Advances*, 9(36):eadh2301, 2023.

[3] Suhrid Ghosh, Anna Koerte, Giulia Serafini, Vinca Yadav, and Jonathan Rodenfels. Developmental energetics: Energy expenditure, budgets and metabolism during animal embryogenesis. In *Seminars in cell & developmental biology*, volume 138, pages 83–93. Elsevier, 2023.

[4] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[5] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[9] Mehmet F Keleş, Ali Osman Berk Sapci, Casey Brody, Isabelle Palmer, Anuradha Mehta, Shahin Ahmadi, Christin Le, Öznur Taştan, Sündüz Keleş, and Mark N Wu. Flyvista, an integrated machine learning platform for deep phenotyping of sleep in drosophila. *Science Advances*, 11(11):eadq8131, 2025.

[10] Cory Pfeiffenberger, Bridget C Lear, Kevin P Keegan, and Ravi Allada. Locomotor activity level monitoring using the drosophila activity monitoring (dam) system. *Cold Spring Harbor Protocols*, 2010(11):pdb–prot5518, 2010.

[11] Jimmy J Fraigne, Jeffrey Wang, Hanhee Lee, Russell Luke, Sara K Pintwala, and John H Peever. A novel machine learning system for identifying sleep–wake states in mice. *Sleep*, 46(6):zsad101, 2023.

[12] Brian Geuther, Mandy Chen, Raymond J Galante, Owen Han, Jie Lian, Joshy George, Allan I Pack, and Vivek Kumar. High-throughput visual assessment of sleep stages in mice using machine learning. *Sleep*, 45(2):zsab260, 2022.

[13] François Laurent, Alexandre Blanc, Lilly May, Lautaro Gándara, Benjamin T Cocanougher, Benjamin MW Jones, Peter Hague, Chloé Barré, Christian L Vestergaard, Justin Crocker, et al. Larvatagger: manual and automatic tagging of drosophila larval behaviour. *Bioinformatics*, 40(7):btae441, 2024.

[14] Jean-Yves Tinevez, Nick Perry, Johannes Schindelin, Genevieve M Hoopes, Gregory D Reynolds, Emmanuel Laplantine, Sebastian Y Bednarek, Spencer L Shorte, and Kevin W Eliceiri. Trackmate: An open and extensible platform for single-particle tracking. *Methods*, 115:80–90, 2017.

[15] Michael Thane, Emmanouil Paisios, Torsten Stöter, Anna-Rosa Krüger, Sebastian Gläß, Anne-Kristin Dahse, Nicole Scholz, Bertram Gerber, Dirk J Lehmann, and Michael Schleyer. High-resolution analysis of individual drosophila melanogaster larvae uncovers individual variability in locomotion and its neurogenetic modulation. *Open Biology*, 13(4):220308, 2023.

[16] Milan Szuperak, Matthew A Churgin, Austin J Borja, David M Raizen, Christopher Fang-Yen, and Matthew S Kayser. A sleep state in drosophila larvae required for neural stem cell proliferation. *Elife*, 7: e33220, 2018.

[17] Parag K Bhatt and Wendi S Neckameyer. Functional analysis of the larval feeding circuit in drosophila. *Journal of visualized experiments: JoVE*, (81):51062, 2013.

[18] Amanda Lesar, Javan Tahir, Jason Wolk, and Marc Gershow. Switch-like and persistent memory formation in individual drosophila larvae. *Elife*, 10:e70317, 2021.

A Detailed methodology

A.1 Animals and video acquisition

Adult Drosophila were maintained on standard molasses-based diet (8.0% molasses, 0.55% agar, 0.2% Tegosept, 0.5% propionic acid) at 25°C on a 12:12 light:dark (LD) cycle. In order to collect synchronized third instar larvae, adult flies were placed in an embryo collection cage (Genesee Scientific, cat#: 59-100) and eggs were laid on a petri dish containing a molasses-based diet with yeast paste on top. Animals developed on this media for three days. After this time, molting 3rd instar Drosophila larvae were manually selected and placed into individual wells of a modified LarvaLodge containing 95 µl of 3% agar and 2% sucrose media covered with a thin layer of yeast paste. The LarvaLodge was covered with a transparent acrylic sheet and placed under a camera for recording. Videos were acquired in 5 minute intervals in IC Capture (The Imaging Source) using an Imaging Source DMK 23UP031 camera (2592 X 1944 pixels, The Imaging Source, USA) equipped with a Fujinon lens (HF12.55A-1, 1:1.4/12.5 mm, Fujifilm Corp., Japan) and a Hoya 49 mm R72 Infrared Filter.

A.2 Manual labeling

Body parts labeling using SLEAP: Anatomical annotation was performed using SLEAP's graphical user interface. Video recordings were loaded into the SLEAP software, and four key anatomical landmarks were manually identified and marked: head, neck, center (torso), and tail. To define the animal's skeletal structure, three edges were established connecting these landmarks: head-neck, neck-center, and center-tail, creating a simplified but biologically meaningful representation of the animal's body configuration. For each video in the dataset, 20 equispaced frames (1 to 2% of total frames) were selected for manual annotation to ensure temporal coverage and provide sufficient training data for the pose estimation model.

Behavior labeling by domain expert: Behavioral annotation was conducted by domain experts among the authors who examined each video frame-by-frame and classified animal behaviors into three primary behavioral states: *sleeping* (periods of inactivity with no body position movement), *feeding* (periods of only mouth part movement), and *crawling* (locomotory behavior characterized by coordinated crawling behavior). Each behavioral epoch was carefully delineated with precise start and end frame indices. Frames that did not clearly exhibit any of the three primary behaviors, or contained ambiguous behavioral states, were assigned the label 'other' in order to accomplish complete timeline coverage. Later, 'other' frames, due to their small number, were excluded from analysis. Behavior labels were stored as Python dictionaries mapping frame ranges to behavioral states for downstream processing.

A.3 Pose estimation using SLEAP

Following the manual annotation of anatomical body parts, a SLEAP pose estimation model was trained to automatically identify (and optionally track) the four body parts across all frames of all videos.

Model architecture and training: The pose estimation pipeline employed a two-stage 'multi-animal top-down' approach that uses pretrained models optimized for multi-animal pose estimation. It consists of two sequential deep learning models: The first, a **centroid detection model**, to identify animal instances within each frame and the second, a centered **instance model**, for precise landmark localization. Both models utilize a U-Net backbone architecture. A variety of parameter values were tried with mixed success in terms of validation test accuracy and quality of prediction on unlabeled frames (as judged by visual inspection by a domain expert), before settling on the values shown in Table 2. Users are recommended to follow a similar process for their datasets: unfortunately, it is not possible to determine these parameters beforehand.

Pose prediction: The trained models were used to generate pose landmark predictions for every frame of every video. The centroid model first identifies animal instances, followed by the centered instance model for precise landmark coordinates. Optionally, a cross-frame tracking algorithm (called 'simple' and packaged with SLEAP) can be used to establish the identity and trajectory of each animal. This inference process produces structured analysis files (`.csv` format) containing time-series data of `(x, y)` coordinates for each landmark for each animal in each video frame, along with confidence scores for each prediction.

A.4 Feature engineering and pose normalization

Raw coordinate data from SLEAP was transformed through comprehensive feature engineering to create meaningful behavioral descriptors. The 12-feature set includes:

Center-normalized coordinates (6 features): All pose coordinates were normalized relative to the animal's *center* landmark to achieve position invariance: $head\_rel\_x = head.x - center.x$, $head\_rel\_y = head.y - center.y$, with similar transformations applied to *neck* and *tail* coordinates.

Table 2: SLEAP parameters for training of Centroid and Centered-Instance models

Parameter	Centroid Model	Centered-Instance
Data		
Validation fraction	0.1	0.1
Input Scaling	0.5	1
Crop size	0 (set by Auto)	320 (set by Auto)
Optimization		
Batch size	16	16
Epochs	30	30
Initial learning rate	1e-04	1e-04
Stop on plateau	True	True
Plateau min. delta	1e-08	1e-08
Plateau patience	20	10
Augmentation		
Rotation	True	True
Rotation min angle	-15	-15
Rotation max angle	15	15
Model		
Backbone	U-Net	U-Net
Max stride	16	32
Filters	16	24
Filters rate	2	2
Middle block	True	True
Up interpolate	True	True
Heads	centroid	centered_instance
Anchor part	center	center
Sigma	2.5	2.5
Output stride	2	4
Size (after training)	30 MB	268 MB

Inter-landmark distances (4 features): Four key distances were calculated to capture postural configurations: head-neck distance (captures head orientation and movement), head-center distance (indicates overall body extension), neck-center distance (torso configuration), and center-tail distance (reflects body posture and tail position).

Body curvature (1 feature): Computed as the absolute angle between head-center and center-tail vectors:

$$curvature = |\arccos(\frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}||\vec{v_2}|})|$$

where,

$\vec{v_1} = (head.x - center.x, head.y - center.y)$ and

$\vec{v_2} = (tail.x - center.x, tail.y - center.y)$

Activity level (1 feature): Velocity-based movement measure computed as the rolling 10-frame average of center-of-mass displacement between consecutive frames, providing a temporal smoothing of instantaneous movement.

All features were standardized using StandardScaler to ensure comparable scales across different feature types, with the exception of body curvature which was already normalized to $[0, \pi]$ radians.

A.5 Validation strategy

Multiple data splitting strategies were evaluated to address the unique challenges of temporal behavioral data:

Stratified random frame sampling - random selection of 20% of frames from each behavior class, maintaining class balance but potentially inflating performance due to temporal correlations between nearby frames.

Temporal block-based split - dividing video data into non-overlapping 15-frame chunks (approximating duration of 1 second) and allocating entire blocks to train (80.9%) or test (19.1%) sets, ensuring temporal independence while preserving behavioral context within blocks.

Table 3: Train-Test split characteristics by validation strategy

Split Strategy	Entity	Train Data	Test Data	Sleep Train/Test	Feed Train/Test	Crawl Train/Test
Stratified Random	Frames	13,241	3,311	3,054/763	7,332/1,833	2,855/715
Temporal Blocks	Blocks	831	195	188/40	472/110	171/45
	Frames	12,417	2,905	2,885/666	7,020/1,654	2,512/585
Animal-Level	Animals	11	2	8/1	9/2	9/1
	Frames	13,640	2,912	3,460/357	6,848/2,317	3,332/238

Animal-level split - entire animals allocated to training or testing sets to evaluate cross-individual generalization, where two animals with distinct behavioral profiles (one pure feeding animal: 1647 frames, 100% feeding; one mixed behavior animal: 1265 frames, 53% feeding, 28% sleeping, 19% crawling) serve as test subjects, constituting 17.6% of the dataset.

The temporal block split was specifically developed to address the fundamental challenge of achieving both temporal independence and balanced class representation in behavioral time series data. By processing only uniform behavior blocks (containing a single behavior class throughout the 15-frame window), this strategy ensures clean behavioral labels while reducing the risk of overfitting to temporal autocorrelations. This filtering removes mixed-behavior blocks. Table 3 summarizes the train-test distribution characteristics for each strategy.

A.6 Machine learning model specifications

Five machine learning classifiers were implemented with conservative hyperparameters. All models used `random_state=42`.

Random Forest Classifier - 50 trees chosen for balance between performance and computational efficiency on the 16K frame dataset.

Support Vector Machine - radial basis function kernel selected for non-linear behavioral patterns.

Gradient Boosting Classifier - 50 boosting stages to prevent overfitting while maintaining learning capacity.

K-Nearest Neighbors - five neighbors chosen to balance local sensitivity with noise robustness.

Neural Network - Two hidden layers consisting of 12 and 6 neurons with early stopping for convergence.

Table 4: Dataset composition before and after processing and cleaning

Dataset Characteristic	Value
Raw Data	
Initial frames (all videos)	18,483
Labeled frames (SLEAP)	13,714
Videos analyzed	10
Data Processing	
Frames with missing coordinates	205
Frames removed (poor visibility)	1,726
Final clean dataset	16,552
Final Dataset Composition	
Total animals tracked	13
Feeding frames	9,165 (55.4%)
Sleeping frames	3,817 (23.1%)
Crawling frames	3,570 (21.6%)
Feature Engineering	
Original pose coordinates	8
Relative pose coordinates	6
Additional engineered features	6

A.7 Behavior evaluation

Model performance was assessed using standard classification metrics computed on a per-frame basis: accuracy (overall correct predictions), precision (true positives per predicted class), recall (true positives per actual class), and F1-score (harmonic mean of precision and recall). These metrics were calculated on the test set separately for each validation strategy. Confusion matrices were generated for the best-performing model in each validation strategy to provide detailed insight into class-specific performance patterns and common misclassification behaviors. Frame-by-frame temporal visualizations were created to enable qualitative assessment of prediction consistency and identification of systematic errors across different validation strategies.

Table 5: Individual animal behavioral profiles

Animal	**Total Frames**	**Sleeping**	**Feeding**	**Crawling**
video 1 animal 2	1,647	0	1,647	0
video 2 animal 1	1,279	0	1,192	87
video 2 animal 2	1,265	357	670	238
video 3 animal 1	1,025	954	0	71
video 3 animal 2	980	8	194	778
video 4 animal 1	911	459	321	131
video 5 animal 1	771	724	0	47
video 5 animal 2	898	66	184	648
video 6 animal 1	1,455	32	1,004	419
video 7 animal 1	1,588	428	1,160	0
video 8 animal 1	1,588	588	1,000	0
video 9 animal 1	1,571	0	1,234	337
video 10 animal 1	1,574	201	559	814
Total	16,552	3,817	9,165	3,570

A.8 Computational resources

Manual pose and behavior labeling was performed using SLEAP GUI and LosslessCut on various consumer grade laptops and PCs. A High Performance Computing Center node powered by two Intel Xeon Gold 6130 32 core CPUs and one A100 GPU running Centos 6.5 was used for running SLEAP GUI for pose training and inference. Training took approximately 3 minutes and 1.5 minutes per epoch for the centroid and the centered instance models respectively. Inference on all frames of all videos took around 30 minutes. A Hyperplane 8 server powered by two Intel Xeon Gold 6248 20 core CPUs and 8 Tesla V100 GPUs running Ubuntu 20.04 was used for behavior classification tasks in Python. Training time for behavior classification models ranged from seconds to minutes depending on the number of features, training set size, and model parameters.

B Dataset composition and processing results

The dataset demonstrates substantial behavioral diversity across individual animals, with three distinct behavioral profiles emerging: feeding-dominant animals (7/13), sleeping-dominant animals (3/13), and crawling-dominant animals (3/13) (Tables 4 and 5). This individual variation provides diverse training data for behavioral classification models while ensuring generalizability across different behavioral phenotypes.

C Behavior timeline and confusion matrices

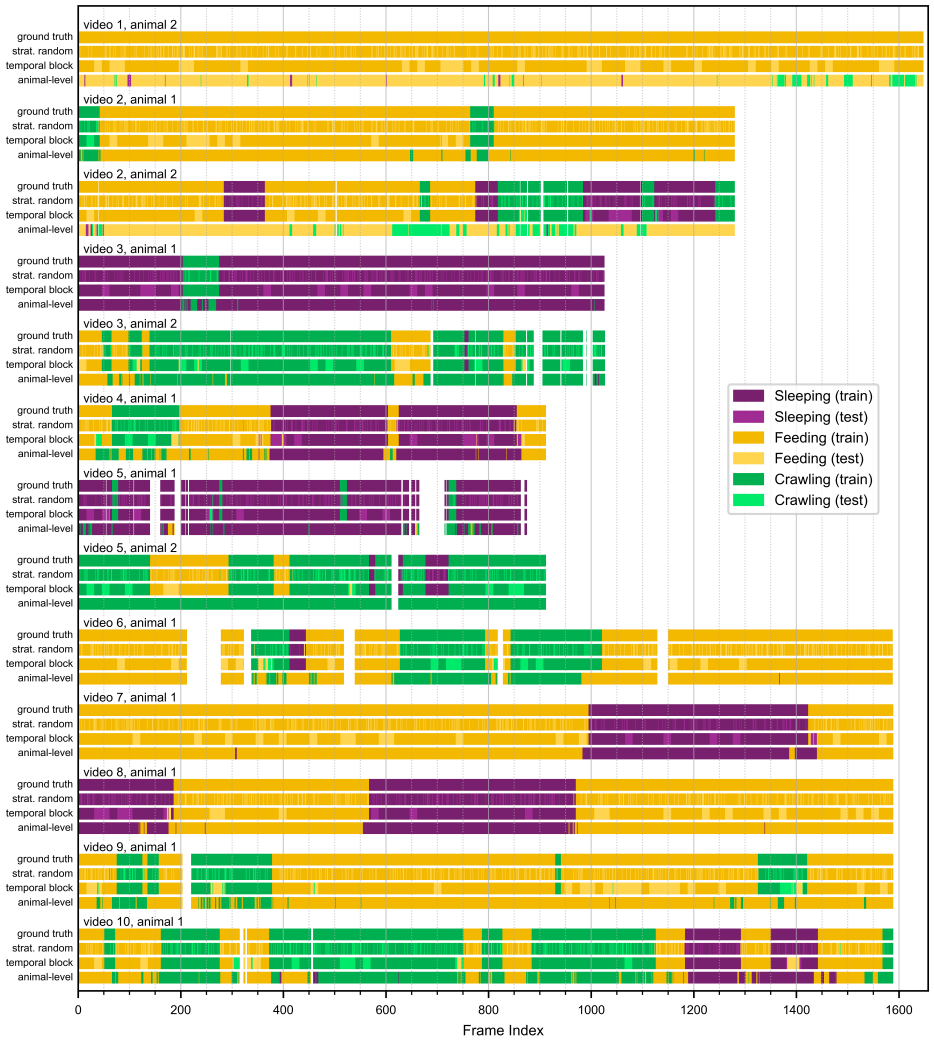

Figure 2: Behavior timeline for each animal showing the ground truth labels for behavior as the top bar and the predictions by the best performing model for each of the three validation strategies.

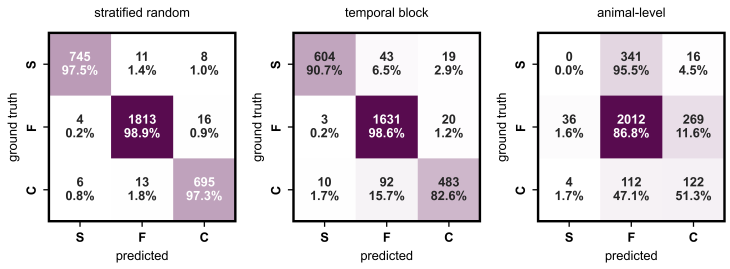

Figure 3: Confusion matrices for the best performing models for each validation strategy.