Automated behavioral annotation in rodent trials using AI frameworks

by

Emily L Alfs Votipka

B.S, Doane University, 2016

M.S, Kansas State University, 2018

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Abstract

Behavioral assays in rodents are fundamental tools for studying cognition, anxiety, and neurological disorders, yet manual annotation of animal behaviors remains labor-intensive, subjective, and difficult to scale. This dissertation presents an automated framework for behavioral annotation in rodent trials that integrates deep learning, pose estimation, and rule-based spatial reasoning to produce accurate, reproducible, and analysis-ready behavioral data from video recordings. The framework is developed primarily for the Novel Object Recognition Test (NORT) and is further extended to the Elevated Plus Maze (EPM) and Marble Burying Test (MBT), demonstrating its adaptability across multiple behavioral paradigms.

The proposed workflow combines fine-tuned YOLOv11 pose estimation for rat keypoint detection, heuristic-based behavioral inference using spatial relationships between animals and experimental objects, and YOLOv11 image classification models for frame-level behavior recognition. Keypoint models were trained to detect the rat nose and tail-base, enabling estimation of body orientation, posture, and object proximity. Multiple behavior-classification strategies, including heuristic-based, deep learning-based, and hybrid approaches, were evaluated for automated identification of behaviors such as object interaction, standing, freezing, resting, and marble interaction. The resulting framework also generates analysis-ready outputs, including behavioral summaries, heatmaps, ethograms, and quantitative metrics for downstream neuroscience analyses.

A central feature of the framework is the separation of generic learned behavioral recognition from task-specific spatial interpretation, enabling visual information to be shared across related experimental configurations while retaining explicit object- and region-specific behavioral outputs.

Experimental evaluation demonstrated that a hybrid approach combining a unified YOLOv11 behavior classification model with keypoint-based heuristic refinement achieved the best overall performance across two-object and five-object NORT configurations while generalizing effectively to varying experimental conditions. The framework was subsequently applied to over 500 NORT videos to enable large-scale analysis of object exploration, novel object and place recognition, rearing, and resting/freezing behaviors across experimental groups and testing sessions. Additional adaptations for different strains of rats, EPM analysis, and MBT further demonstrate the flexibility of the proposed methodology for diverse rodent behavioral assays.

This work provides a unified, scalable framework for automated rodent behavior annotation that substantially reduces manual scoring effort while maintaining high analytical utility. By integrating modern computer vision techniques with interpretable spatial reasoning, the proposed approach advances automated behavioral neuroscience and establishes a flexible foundation for future studies involving additional behavioral paradigms, animal strains, and experimental environments.

Automated behavioral annotation in rodent trials using AI frameworks

by

Emily L Alfs Votipka

B.S, Doane University, 2016

M.S, Kansas State University, 2018

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Doina Caragea

Copyright

© Emily L Alfs Votipka 2026.

Abstract

Behavioral assays in rodents are fundamental tools for studying cognition, anxiety, and neurological disorders, yet manual annotation of animal behaviors remains labor-intensive, subjective, and difficult to scale. This dissertation presents an automated framework for behavioral annotation in rodent trials that integrates deep learning, pose estimation, and rule-based spatial reasoning to produce accurate, reproducible, and analysis-ready behavioral data from video recordings. The framework is developed primarily for the Novel Object Recognition Test (NORT) and is further extended to the Elevated Plus Maze (EPM) and Marble Burying Test (MBT), demonstrating its adaptability across multiple behavioral paradigms.

The proposed workflow combines fine-tuned YOLOv11 pose estimation for rat keypoint detection, heuristic-based behavioral inference using spatial relationships between animals and experimental objects, and YOLOv11 image classification models for frame-level behavior recognition. Keypoint models were trained to detect the rat nose and tail-base, enabling estimation of body orientation, posture, and object proximity. Multiple behavior-classification strategies, including heuristic-based, deep learning-based, and hybrid approaches, were evaluated for automated identification of behaviors such as object interaction, standing, freezing, resting, and marble interaction. The resulting framework also generates analysis-ready outputs, including behavioral summaries, heatmaps, ethograms, and quantitative metrics for downstream neuroscience analyses.

A central feature of the framework is the separation of generic learned behavioral recognition from task-specific spatial interpretation, enabling visual information to be shared across related experimental configurations while retaining explicit object- and region-specific behavioral outputs.

Experimental evaluation demonstrated that a hybrid approach combining a unified YOLOv11 behavior classification model with keypoint-based heuristic refinement achieved the best overall performance across two-object and five-object NORT configurations while generalizing effectively to varying experimental conditions. The framework was subsequently applied to over 500 NORT videos to enable large-scale analysis of object exploration, novel object and place recognition, rearing, and resting/freezing behaviors across experimental groups and testing sessions. Additional adaptations for different strains of rats, EPM analysis, and MBT further demonstrate the flexibility of the proposed methodology for diverse rodent behavioral assays.

This work provides a unified, scalable framework for automated rodent behavior annotation that substantially reduces manual scoring effort while maintaining high analytical utility. By integrating modern computer vision techniques with interpretable spatial reasoning, the proposed approach advances automated behavioral neuroscience and establishes a flexible foundation for future studies involving additional behavioral paradigms, animal strains, and experimental environments.

Table of Contents

List of Figures xi

List of Tables xiii

List of Algorithms xv

Acknowledgments xvi

Dedication xvii

1 Introduction 1

2 Overview of YOLOv11 17

3 Video data collection 20

3.1 NORT videos 21

3.2 EPM videos 23

3.3 MBT videos 24

4 Keypoint annotation 25

4.1 YOLOv11 pose model for rat keypoint annotation 25

4.2 Manual frame-based keypoint annotation 27

4.3 Metrics for keypoint annotation 29

4.4 YOLOv11 keypoint annotation model performance 30

5 Novel object recognition test (NORT) 35

5.1 NORT behaviors of interest 35

5.2 Manual frame-level behavior annotation . . . 36
5.3 Keypoint-based heuristics for frame-level behavior classification . . . 41
5.4 YOLOv11 models for frame-level behavior classification . . . 43
5.5 Evaluation metrics for NORT behavior classification . . . 45

6 Validation-based optimization for NORT models . . . 47
6.1 Hyperparameter optimization for the heuristic-based approach . . . 47
6.2 Hyperparameter optimization for YOLOv11 classification models . . . 51
6.3 Post-processing optimization . . . 52
6.4 Human agreement comparison . . . 55

7 NORT results . . . 59
7.1 Performance results for three-class predictions . . . 59
7.2 Performance results for specific object interactions . . . 60
7.3 Error analysis for behavior classification . . . 62

8 Analysis-ready outputs for NORT . . . 68
8.1 Final output files . . . 68
8.2 Visualization outputs . . . 70
8.2.1 Heatmaps . . . 70
8.2.2 Ethograms . . . 73

9 Framework extensions for Fisher 344 NORT and EPM analysis . . . 76
9.1 Overview of framework extensions . . . 76
9.2 Adaptation to Fisher 344 NORT videos . . . 77
9.2.1 Keypoint model adaptation . . . 78
9.2.2 Behavior classification adaptation . . . 80
9.2.3 Discussion of Fisher 344 adaptation . . . 83
9.3 Elevated Plus Maze analysis . . . 85

9.3.1 EPM keypoint adaptation and preprocessing . . . 85
9.3.2 EPM region definition . . . 87
9.3.3 Frame-level region assignment and bout counting . . . 88
9.3.4 EPM summary outputs and visualizations . . . 89

10 Marble burying and marble-interaction analysis . . . 93
10.1 Overview and behavioral objectives . . . 93
10.2 Rat keypoint adaptation for MBT videos . . . 95
10.3 Marble segmentation and identity tracking . . . 98
10.4 Heuristic frame-level marble-interaction labeling . . . 105
10.5 Manual behavior annotation . . . 111
10.6 YOLOv11 behavior classification for MBT . . . 112
10.7 Preliminary discussion and limitations . . . 115

11 Conclusions and future work . . . 117
11.1 Conclusions . . . 117
11.2 Future work . . . 120

Bibliography . . . 123

A Global NORT video analysis . . . 128

List of Figures

1.1 NORT arenas 4

1.2 EPM arena 5

1.3 MBT arena 6

1.4 Workflow diagram 11

2.1 General YOLOv11 architecture 19

3.1 Long-Evans rat 20

3.2 Fisher 344 rat 21

3.3 NORT setup for 2-object and 5-object videos 22

4.1 LableMe interface 27

4.2 Challenging poses for keypoint annotation 28

4.3 Duration of sequential missing keypoints. 31

4.4 Nose annotation analysis post-interpolation 32

4.5 Sample interpolation results for the *nose* keypoint 33

5.1 CVAT interface 36

5.2 Difficult behavior transition frames 38

6.1 Arena bounding boxes 48

6.2 Bounding areas for objects 50

6.3 Confusion matrices comparing student and model 58

7.1 Confusion matrices of YOLOv11-`large` on the NORT behavior prediction . . 64

7.2 Confusion matrices for YOLOv11-`medium-post` 65

7.3 Confusion matrices for heuristic-based approach 66
7.4 Behavior annotation error analysis 67

8.1 Sample heatmaps 71
8.2 Radiating of binned points 72
8.3 Sample ethogram for 2-object NORT 75

9.1 Modified keypoint model on Fisher 344 79
9.2 Comparison of Long-Evans model and fine-tuned Fisher 344 models 81
9.3 Errors present in Fisher 344 with fine-tuned model 82
9.4 Modified keypoint model on EPM 86
9.5 EPM region definition procedure 88
9.6 EPM heatmap 90
9.7 EPM ethogram 91

10.1 Nose keypoint verification for MBT videos 97
10.2 Comparison of Long-Evans model and fine-tuned MBT models 99
10.3 Errors present in MBT with fine-tuned model 100
10.4 Example marble predictions 102
10.5 Marble identification 104
10.6 Heuristic procedure for marble-interaction labeling 109
10.7 Arena standing boundary 110

A.1 Behavior of rats on days 3 and 4 129
A.2 Difference in time spent in position 4 on day 3 and day 2 130
A.3 Difference in time spent in position 1 on day 4 and day 2 131

List of Tables

4.1 Keypoint annotated dataset 29

4.2 Missing keypoint annotations 31

5.1 Behavior annotated dataset 39

5.2 Annotated dataset (`Upper Object`/`Lower Object`) 40

5.3 Annotated behavior dataset (`Object 1` to `Object 5`) 40

6.1 Validation performance of the heuristic approach for the `Standing` behavior 49

6.2 Validation performance of the `Object Interaction` heuristic 50

6.3 Validation performance of varying size YOLOv11 classification models 52

6.4 Validation performance of YOLOv11 classification models 53

6.5 Post-processing validation performance of the YOLOv11-`large` 54

6.6 Post-processing validation performance of the YOLOv11-`medium` 55

6.7 Human agreement results versus combined YOLOv11-`large` model, B8.13_G3 56

6.8 Human agreement results versus combined YOLOv11-`large` model, B6.04_B3 57

7.1 Performance results on test videos 60

7.2 Performance results for specific object interactions in the 2-object test videos 61

7.3 Performance results for specific object interactions in the 5-object test videos 62

8.1 Analysis-ready CSV files generated by the NORT post-analysis pipeline . . . 69

9.1 Expanded keypoint annotation dataset used for Fisher 344 NORT adaptation 78

9.2 Fisher 344 keypoint model performance before and after fine-tuning 79

9.3 Expanded behavior annotation dataset used for Fisher 344 NORT adaptation 82

9.4 Behavior classification performance before and after Fisher 344 fine-tuning . 83

10.1 Expanded keypoint annotation dataset for MBT adaptation 96

10.2 MBT keypoint model performance before and after fine-tuning 97

10.3 Marble segmentation dataset 102

10.4 Heuristic rule value selection 108

10.5 MBT behavior annotation dataset 112

10.6 Validation performance of YOLOv11 model sizes for MBT behavior classification 113

10.7 Validation performance of the selected YOLOv11 MBT behavior classification model 114

10.8 Class-level test performance for MBT behavior classification 115

List of Algorithms

8.1 Heatmap generation algorithm 72

8.2 Ethogram generation algorithm 74

10.1 Marble tracking algorithm 106

Acknowledgments

I would like to express my deepest gratitude to my major professor, Dr. Doina Caragea, whose guidance, mentorship, and unwavering support have been instrumental throughout my doctoral journey. Your generosity with your time, thoughtful advice, and steadfast encouragement challenged me to become a better scholar and gave me the confidence to persevere through every stage of this work.

I am sincerely grateful to the esteemed members of my doctoral committee for their commitment throughout this journey. Your expertise and encouragement have strengthened both this dissertation and my development as a scholar.

To my many collaborators, thank you for your invaluable contributions, insightful discussions, and willingness to share your knowledge. Your partnership enriched this work in countless ways, and I am grateful for the opportunity to have learned alongside you.

Finally, to my dear friends and family, thank you for your unwavering encouragement and steadfast belief in me. Your love and support have sustained me through every challenge and made this achievement all the more meaningful.

This research was also partially sponsored by the Cognitive and Neurobiological Approaches to Plasticity (CNAP) Center of Biomedical Research Excellence (COBRE) of the National Institutes of Health (NIH) under grant number P20GM113109. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Dedication

To my mother, whose unwavering belief in me never faltered, even when my own confidence did. You saw possibilities in me long before I could see them for myself, and your love, encouragement, and faith became a constant source of strength throughout my life.

Your presence has been with me every step of this journey. The lessons you taught me, the confidence you instilled in me, and the love you gave continue to guide me every day.

This dissertation is dedicated to your memory, with profound love, gratitude, and the hope that I have honored the faith you always had in me. Though you are no longer here to witness its completion, it is your footsteps that led me to this finish line.

Chapter 1

Introduction

In neuroscience and behavioral research, the task of accurately identifying and quantifying specific behaviors is essential for understanding underlying biological mechanisms relevant to cognitive health (Schlinger, 2015). Rat assays have been widely used for this task in both basic and translational research, as rats exhibit a rich and diverse repertoire of behavioral phenotypes relevant to cognitive and disease-related processes that are crucial for studying a range of phenomena, from cognitive processes to disease mechanisms (Powers, 1971). Traditional approaches to animal behavior identification have heavily relied on manual coding. However, manual coding requires significant human effort, despite the availability of tools that aid the coding process, and is often prone to subjectivity and potential biases in data interpretation, given that animal behaviors can occur in complex and sometimes subtle patterns. Furthermore, given the increasing volume of behavioral data generated by high-throughput experimental paradigms, e.g., automated video tracking systems, manual coding has limited scalability and becomes unsustainable, especially when large datasets are required for robust statistical analysis (Anderson and Perona, 2014).

The need for efficient, objective, and scalable methods of animal behavior annotation has led to growing interest in automated solutions. Advances in machine learning, computer vision, and deep learning offer the potential to significantly enhance the accuracy, efficiency, and consistency of behavior annotation (Fazzari et al., 2024; van Dam et al., 2020; von Ziegler

et al., 2021). By leveraging these technologies, it is possible to accelerate data processing, reduce human error, and provide reproducible results, which are crucial for the credibility and reliability of scientific findings.

Various machine learning-based approaches have been developed in recent years to enable automated animal behavior. Some approaches focus on video frame classification using deep neural networks, such as convolutional neural networks (CNNs) (Perez and Toler-Franklin, 2023) to identify more static behaviors (e.g., object interaction or standing behaviors). Other approaches focus on frame sequence classification using recurrent neural networks (RNNs) (Van Dam et al., 2023) or temporal convolutional networks (TCNs) (Whiteway et al., 2021) that aggregate a sequence of video frames to classify dynamic behaviors (e.g., resting or grooming). Keypoint detection is another powerful technique, in which deep learning models, such as DeepLabCut (Mathis et al., 2018) or DeepPoseKit (Graving et al., 2019), are trained to identify and track specific body parts (e.g., head or tail) in animals. The detected keypoints can be used as inputs to machine learning models, which infer specific behaviors, e.g., SimBA (Goodwin et al., 2024). Alternatively, heuristics can be applied to refine the identification of behaviors by leveraging keypoint-based rules or patterns derived from biological knowledge, e.g., BehaviorDEPOT (Gabriel et al., 2022) or RAINSTORM (D'hers et al., 2025).

Despite the widespread development of deep learning approaches and models for automated rat behavior identification, the variability in animal genetics and environmental factors make it hard to directly reuse deep learning models across studies. For instance, differences in rat genetics, such as color or size, or environmental factors such as cage design and housing conditions, complicate the use of models for keypoint detection or even rat detection and call for customized models fine-tuned using data from the study at hand (von Ziegler et al., 2021). Similarly, rat behaviors of interest in diverse behavioral neuroscience studies call for models fine-tuned to those specific studies and behaviors (Philipsberg et al., 2023).

Throughout the duration of this research, an assortment of datasets were produced and the prioritization of research tasks was continuously refined to align with the evolving inter-

ests and objectives of collaborating researchers, ensuring that development efforts remained responsive to ongoing experimental demands. The data for this was collected and provided by Dr. Bethany Plakke's lab of Kansas State University. In total, this work consists of three different trial paradigms: Novel Object Recognition Test (NORT), Elevated Plus Maze (EPM), and Marble Burying Test (MBT).

The Novel Object Recognition Test (NORT) paradigm is used to assess learning and memory functions in rodents and focuses on the rodents' innate tendency to explore novel cues (Lueptow, 2017). This strong novelty preference extends not only to objects that were not previously encountered by the rodents, but also to objects that were previously encountered in a different location. The NORT experiment process generally consists of two phases, the sample and test phases, separated by a delay. In the traditional NORT setup, the rodents are allowed to explore two identical objects in the sample phase, one of which is replaced with a novel object in the test phase. If the rodents have intact memory for the previously presented object, they are expected to spend more time exploring the novel object (Lueptow, 2017). The time spent with each object is used by neuroscientists in their analyses. Figure 1.1 shows two frames that illustrate the two types of NORT arenas. The `Object Interaction` behavior is shown in the 2-object frame on the left and the `Standing` behavior is shown in the 5-object video on the right. In the 2-object videos, the objects are referred to by their position in the frame as `Upper` or `Lower` objects as shown in Figure 1.1 (left), while in the 5-object videos, the objects are referred to as `Object 1` to `Object 5`, as shown in Figure 1.1 (right).

The Elevated Plus Maze (EPM) paradigm is a widely used behavioral assay for assessing anxiety-like behavior in rodents. The apparatus consists of four arms arranged in a plus-shaped configuration and elevated above the floor, with two opposing open arms and two opposing closed arms enclosed by walls. The test examines the natural conflict between a rodent's innate drive to explore novel environments and its aversion to open, elevated spaces (Walf and Frye, 2007). Anxiety-like behavior is inferred from the animal's exploration pattern. Rodents typically spend more time in the enclosed arms because these areas provide protection from potential threats. Increased exploration of the open arms, reflected by

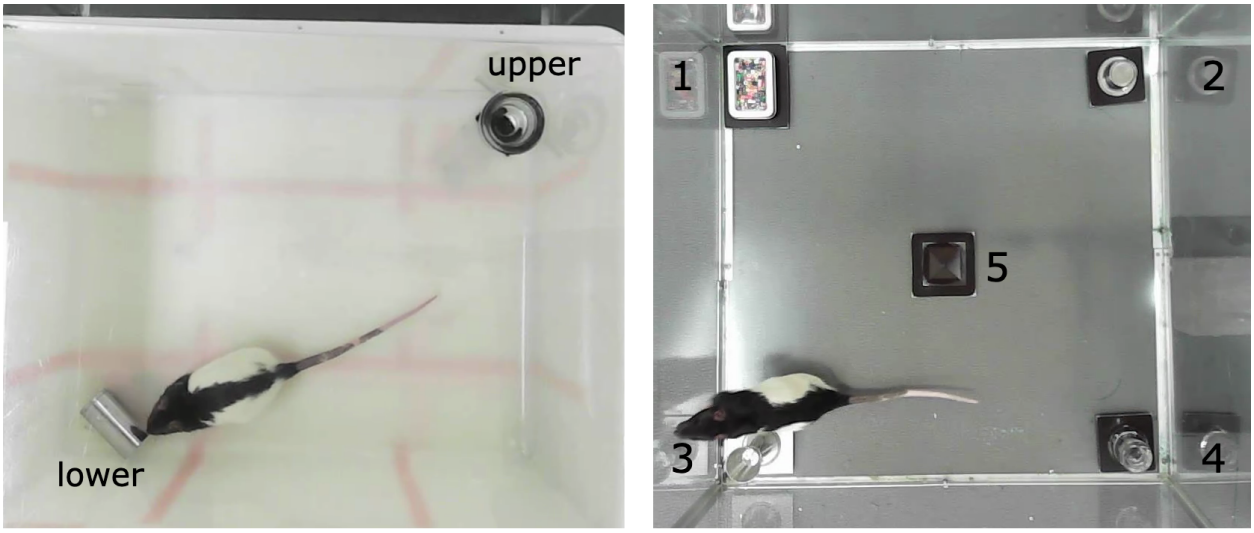


Figure 1.1: *NORT arenas. Frames depicting two types of NORT arenas, specifically 2-object (left) and 5-object (right) arenas. The rat is a Long-Evans rat, a breed characterized by white coat and black hood. The sample frame on the left illustrates the* `Object Interaction` *behavior, while the one on the right illustrates the* `Standing` *behavior. Object naming conventions are also illustrated. Objects are identified by their position in the frame, regardless of what the actual object at that position may be. In the 2-object videos, objects in the upper-right/lower-left corners of the frame are referred to as* `Upper/Lower`*, respectively. In the 5-object videos, the object in the upper-left corner is* `Object 1`*, the object in the upper-right corner is* `Object 2`*, the object in the lower-left corner is* `Object 3`*, the object in the lower-right corner is* `Object 4`*, and the object in the middle is* `Object 5`*.*

greater time spent or more entries into open arms, is generally interpreted as a relaxed response, whereas reduced open-arm exploration indicates increased anxiety-like behavior (Walf and Frye, 2007). In this trial, total arm entries are used for analysis by neuroscientists. Figure 1.2 shows a sample of the elevated plus maze arena divided into its separate regions.

The Marble Burying Test (MBT) paradigm is a behavioral assay commonly used in rodents to assess repetitive, compulsive-like, anxiety-related, and defensive digging behaviors. The test is based on the natural tendency of rodents to dig and bury unfamiliar objects present within their environment. Although originally developed as a model of anxiety-related behavior, contemporary interpretations recognize that marble burying may reflect a complex combination of defensive responses, species-typical digging behavior, and repetitive or perseverative actions depending on the experimental context and pharmacological manipulations. The primary outcome measure is the total number or percentage of marbles

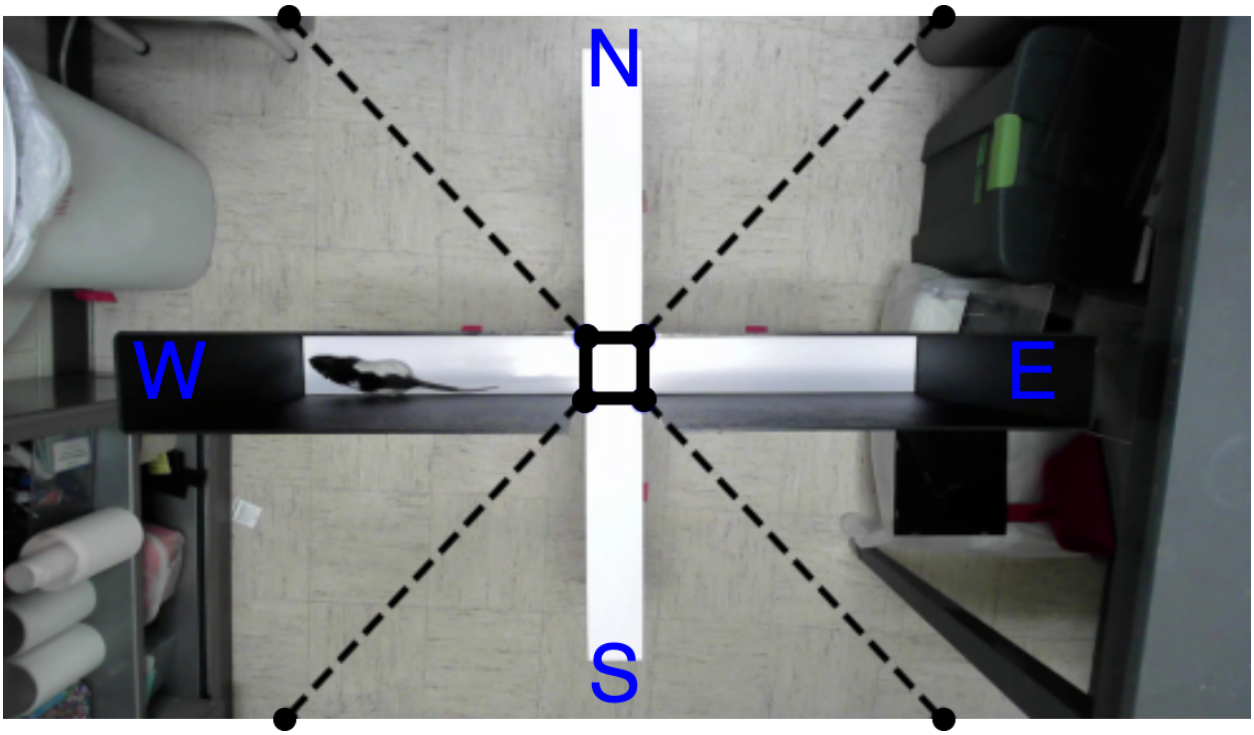


Figure 1.2: *EPM arena. The arena is an elevated plus where the North and South arms are open and exposed. The East and West arms are enclosed offering a perceived safety.*

buried during the testing period. Increased marble burying behavior has been interpreted as reflecting enhanced anxiety-like or compulsive-like responding, whereas reductions in burying behavior are often observed following administration of pharmacologic agents (Deacon, 2006). Figure 1.3 shows the initial configuration for the MBT videos in our dataset.

In each of these paradigms, manually tracking each behavior and the time spent with each object is labor-intensive. For example, manually annotating interactions with two objects in a 10-minute NORT video took up to 40 minutes. This task becomes more challenging when the number of objects is increased, as with variations of the traditional NORT. In the case of MBT, manual annotations for 10 minutes took around 90 minutes due to small size of the marbles as well as the lower video resolution. Thus, manual annotation of the behaviors of interest in a large number of videos is a daunting task and automated approaches are needed.

To address this need, rule-based approaches, such as BehaviorDEPOT (Gabriel et al., 2022), have been designed and used in conjunction with pose/keypoint estimation models.

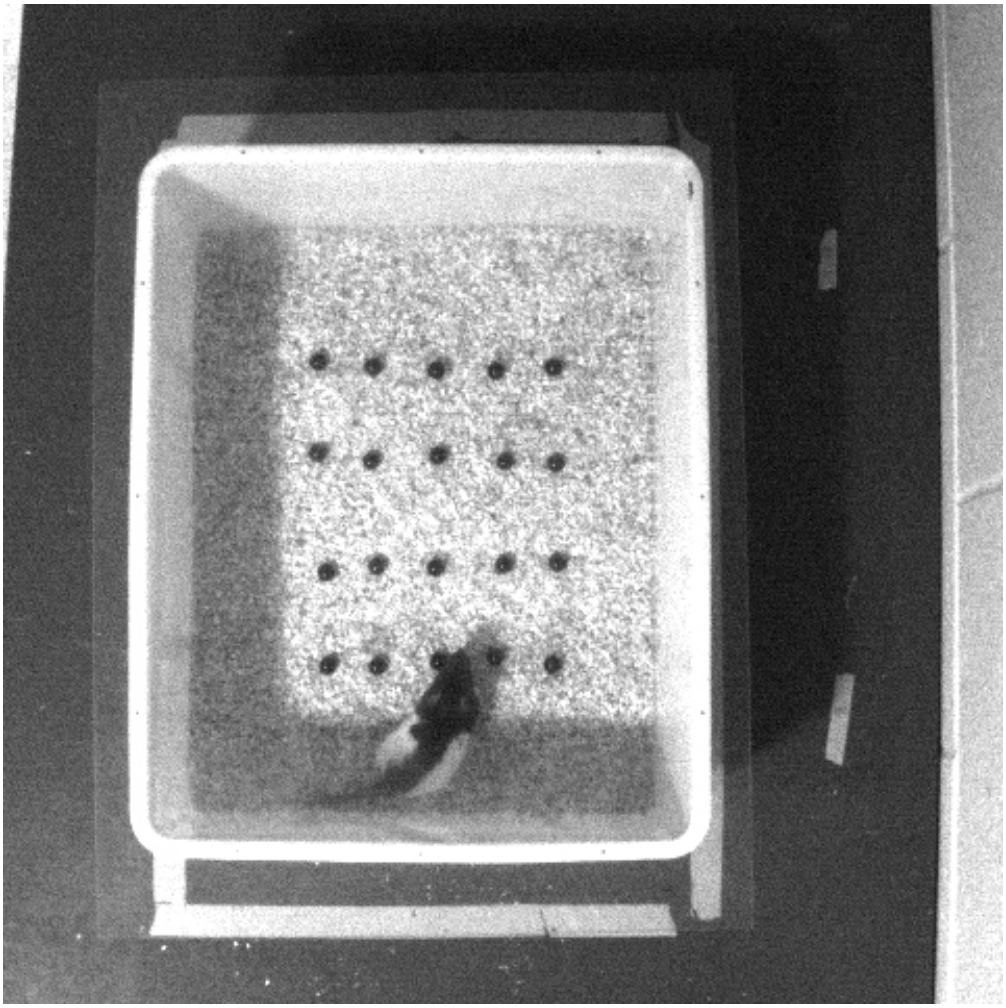


Figure 1.3: *MBT arena. The arena has a layer of bedding and is initialized with an array of 20 marbles evenly spaced.*

These approaches are part of a broader shift toward computational behavioral analysis, in which video recordings are converted into quantitative measures of animal position, pose, and behavior. Markerless pose-estimation frameworks such as DeepLabCut (Mathis et al., 2018), DeepPoseKit (Graving et al., 2019), and SLEAP (Pereira et al., 2022) have made it possible to track experimenter-defined body parts with high spatial and temporal resolution, often with relatively small amounts of manually labeled training data. However, pose estimation alone does not directly produce behavior labels. A second step is typically required to translate trajectories and keypoints into ethologically meaningful behaviors, either through hand-designed rules, supervised classifiers, or end-to-end frame-level models.

Several software systems have addressed this second step. JAABA introduced an interactive machine-learning framework in which users label examples of behaviors and train classifiers from tracking-derived features (Kabra et al., 2013). More recent systems such as SimBA provide graphical interfaces for training supervised behavior classifiers from pose-estimation outputs and include tools for interpreting the learned classifiers (Goodwin et al., 2024). DeepEthogram takes a complementary approach by learning behavior labels directly from raw video pixels, thereby reducing the need for explicit pose annotation in some settings (Bohnslav et al., 2021). BehaviorDEPOT instead emphasizes rule-based behavioral detection from keypoint time series, using kinematic, postural, spatial, and temporal features to define behaviors in a way that is transparent and adaptable to common rodent assays, including open field, elevated plus maze, and novel object exploration (Gabriel et al., 2022). These systems collectively demonstrate the value of automated behavioral annotation, but they also highlight a recurring challenge: behavioral definitions are often experiment-specific and depend on the geometry of the arena, the objects present, the animal strain, and the behaviors of interest.

For object-recognition paradigms such as the Novel Object Recognition Test (NORT), the central challenge is that object exploration is not equivalent to simple proximity. In many studies, exploration is defined by both the animal's position relative to an object and its orientation toward that object. Early work on automated scoring of object recognition in rats demonstrated that object exploration could be quantified computationally, particularly when the animal's nose position could be tracked reliably and when object size and placement were compatible with the tracking assumptions (Rutten et al., 2008). More recently, EXPLORE was developed specifically for object-recognition tests and uses a supervised convolutional neural network to classify whether a rodent is exploring an object based on user-labeled examples (Ibañez et al., 2023). EXPLORE showed that deep learning can reach expert-level performance for object-exploration scoring and can handle complex cases that are difficult for simple point-tracking approaches. Nevertheless, NORT variants with different numbers of objects, arena layouts, object identities, and animal appearances may still require task-specific model adaptation and careful definition of object-specific interactions.

Automated analysis has also been applied to anxiety-related assays such as the Elevated Plus Maze (EPM). Traditional automated tracking systems and open-source tools can estimate locomotion, time spent in open versus closed arms, and arm entries from center-point or body-position tracking. However, EPM behavior can also involve more subtle ethological events, such as head dips, risk assessment, rearing, and transitions at arm boundaries. Sturman et al. (2020) showed that combining DeepLabCut-based pose tracking with supervised machine learning can achieve human-level scoring of ethologically relevant behaviors in commonly used assays, including the EPM, and can outperform commercial solutions for some behavioral measures. Similarly, BehaviorDEPOT provides EPM analysis functions based on keypoint-derived features and spatial regions (Gabriel et al., 2022). These studies support the use of pose-based and machine-learning approaches for anxiety-related behavioral analysis, while also emphasizing that robust performance depends on reliable localization of the animal and accurate mapping between spatial regions and behavioral definitions.

Marble-related assays introduce additional challenges because the behavior of interest may involve both the animal and multiple small objects whose visibility changes over time. The Marble Burying Test (MBT) is widely used in studies of anxiety-like, repetitive, and compulsive-like behaviors, but the interpretation of marble burying remains debated (De Brouwer et al., 2019; Thomas et al., 2009). In the traditional scoring procedure, the primary outcome is the number of marbles buried at the end of the test, which ignores the timing, duration, and sequence of interactions that occur during the trial. Wahl et al. (2022) addressed this limitation by developing an automated approach that uses supervised machine learning to quantify burying bouts and movement characteristics throughout the test. This work shows that video-based analysis can recover richer temporal information than endpoint scoring alone. However, automated marble-interaction analysis remains less developed than general locomotor tracking or NORT object-exploration scoring, particularly when the goal is to detect frame-level interactions between the animal and individual marbles.

Taken together, prior work establishes several complementary strategies for automated rodent behavior analysis. Markerless pose-estimation frameworks provide accurate localization of experimenter-defined body parts but do not themselves assign ethologically meaningful

behavior labels. Supervised and end-to-end classifiers can learn behavior directly from pose-derived features or image data, while rule-based systems provide interpretable behavioral definitions based on animal pose, motion, and experimental geometry. However, these approaches generally emphasize different components of the behavioral-analysis problem, while behavioral definitions often remain specific to the animal, arena, objects, and experimental paradigm.

The gap addressed in this dissertation is therefore not the absence of methods for pose estimation or behavior classification individually, but the need for a modular framework that combines learned visual perception with explicit, experiment-specific spatial reasoning. The proposed framework separates behavioral recognition from the spatial interpretation needed to map a recognized behavior to a particular object, region, or experimental configuration. This separation is developed and rigorously evaluated for NORT and subsequently examined under changes in object configuration, animal appearance, arena geometry, and behavioral objective. In this way, the dissertation investigates not only whether automated behavioral annotation is feasible, but also which components of an annotation pipeline can be shared across experimental settings and which require task- or domain-specific adaptation.

Building on the complementary strengths of interpretable keypoint-based reasoning exemplified by BehaviorDEPOT (Gabriel et al., 2022) and learned visual classification for object exploration exemplified by EXPLORE (Ibañez et al., 2023), we developed and evaluated a hybrid workflow that integrates learned perception with task-specific spatial reasoning for identifying:

- NORT-relevant behaviors, including `Standing`, `Object Interaction`, `Freezing`, and `Resting` behaviors
- EPM-relevant spatial measures, including time spent in each region and arm entries
- MBT-relevant behaviors, including `Standing`, `Marble Interaction`, `Freezing`, and `Resting` behaviors

A high-level overview of the proposed workflow for automated rodent behavior annotation

is shown in Figure 1.4. The figure uses NORT behaviors to illustrate the workflow because NORT is the primary focus of this dissertation and the most extensively developed and evaluated behavioral paradigm in this work. However, the same general framework also informs our extensions to other rodent behavioral tasks, including elevated plus maze (EPM) analysis and marble-interaction analysis. Across these tasks, the workflow combines manual annotation, model-based keypoint detection, task-specific spatial reasoning, and frame-level behavior classification or behavioral summary generation.

As shown in Figure 1.4, we used LabelMe (Russell et al., 2008) to label a set of video frames with rat body keypoints, specifically the *nose* and *tail-base*. These two keypoints provide a compact representation of the animal's position, orientation, and posture. We also used CVAT (CVAT.ai Corporation, 2023) to label a set of NORT frames in terms of task-relevant behaviors, including `Object Interaction` and `Standing`. We first fine-tuned a YOLOv11 Pose model (Khanam and Hussain, 2024) to detect *nose* and *tail-base* keypoints specifically for the Long-Evans rats and arenas used in our videos. Similar to BehaviorDEPOT (Gabriel et al., 2022), we then used these keypoints to design spatially informed heuristic rules for identifying behaviors of interest.

For NORT, the keypoint-based heuristics were used to identify behaviors such as `Object Interaction` and `Standing`. In particular, the predicted *nose* and *tail-base* keypoints allowed us to reason about the rat's proximity to an object, orientation toward the object, and body posture within the arena. These cues are especially important because object exploration in NORT is not simply a matter of being near an object; rather, the animal must be positioned and oriented in a way that indicates active investigation. The NORT task therefore served as the central testbed for developing and evaluating the proposed hybrid workflow.

Furthermore, inspired by EXPLORE (Ibañez et al., 2023), we trained YOLOv11 classification models (Khanam and Hussain, 2024) at frame level independently for the 2-object and 5-object NORT videos, as well as for the combined set of 2-object and 5-object videos. These models were trained to identify `Standing`, `Object Interaction`, and `Other` behaviors. The models trained independently on 2-object or 5-object videos can identify specific object interactions, such as `Upper` or `Lower` in 2-object videos and `Object1` to `Object5` in 5-object

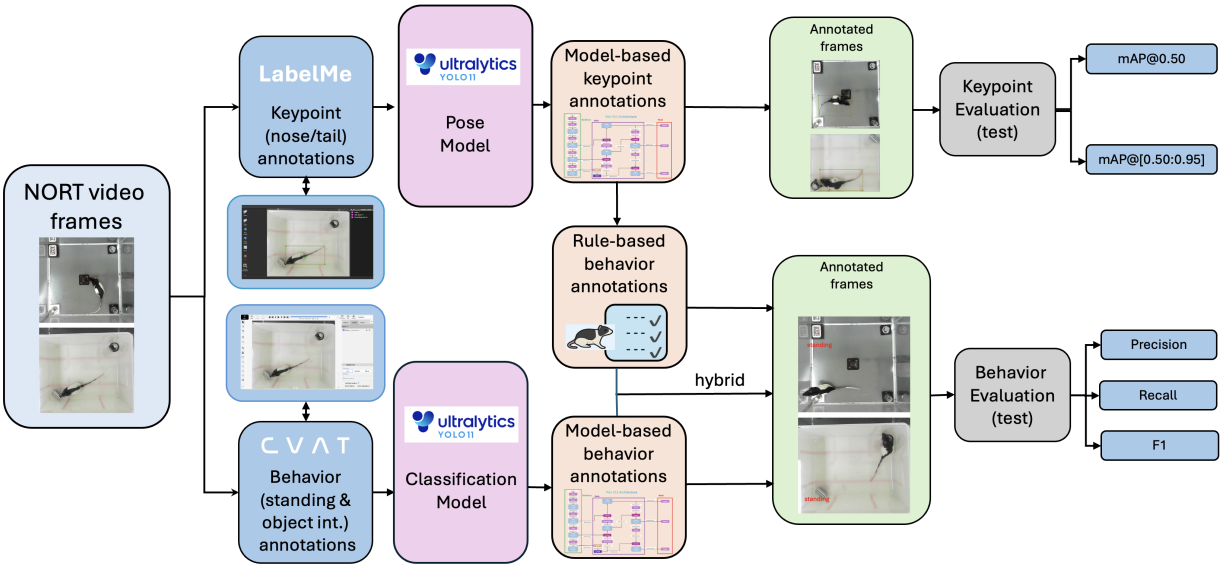


Figure 1.4: *Workflow diagram illustrating the proposed automated annotation framework using NORT as the primary example. Both 2-object and 5-object NORT video frames are labeled in terms of keypoints using LabelMe and task-relevant behaviors using CVAT. YOLOv11 Pose and Classification models are trained for keypoint annotation and frame-level behavior classification, respectively. Predicted keypoints are then used in a heuristic-based approach to identify NORT behaviors and to refine the predictions of the YOLOv11 behavior classification model. Although the figure focuses on NORT, the same general workflow motivates the extensions to EPM and marble-interaction analysis, where keypoints, arena regions, and object locations are adapted to the behavioral outcomes of interest.*

videos. In contrast, the model trained on the combined set identifies generic object interactions. To transform generic object-interaction predictions into specific object interactions, we leveraged the keypoint-based heuristic approach to determine the specific object location, represented by a bounding box or circular region, with which the rat was interacting. Training on the combined 2-object and 5-object videos has the potential to improve robustness and generalization to NORT arenas with different numbers or arrangements of objects. This decomposition is central to the proposed framework: visual recognition of a generic behavioral state is shared across experimental configurations, whereas configuration-specific information is handled explicitly by the spatial-reasoning component. Consequently, the behavior classifier does not need to learn a separate class for every combination of behavior, object identity, and arena configuration.

We compared three NORT-focused approaches: a keypoint-based heuristic approach,

YOLOv11 classification models trained independently on the 2-object and 5-object videos, and a heuristic-enhanced YOLOv11 classification model trained on the combined set of 2-object and 5-object videos. Experimental results showed that the combined model, enhanced with keypoint-based heuristics to identify specific object interactions, achieved the best overall performance. We therefore used this model to perform a large-scale analysis of approximately 500 NORT videos in terms of group differences in object exploration, novel object recognition, rearing behavior, and resting/freezing behavior across multiple days.

Beyond NORT, we also explored how the proposed workflow can be adapted to additional rodent behavioral paradigms. For EPM, the focus shifts from object interaction to spatial localization within predefined regions of the apparatus, such as open arms, closed arms, and the center region. In this setting, keypoint predictions and arena-region definitions can support the extraction of behavioral measures such as arm entries and time spent in each region. For marble-interaction analysis, the workflow is adapted to account for both the animal's body keypoints and the locations of multiple small objects. This extension supports the identification of frames in which the rat is positioned near and oriented toward marbles, providing a foundation for more detailed analysis of marble-directed behaviors. These EPM and marble-interaction components are presented as extensions of the core framework, demonstrating the adaptability of the proposed approach beyond the primary NORT setting.

To summarize, the main contributions of this dissertation are as follows:

- **Hybrid and factorized behavioral annotation.** We developed and evaluated a hybrid framework that combines learned frame-level behavioral recognition with explicit keypoint-based spatial reasoning. In the NORT setting, the framework separates generic recognition of behaviors such as Object Interaction from identification of the specific object involved, allowing the learned classifier to share information across experimental configurations while retaining object-specific behavioral outputs.

- **Cross-configuration generalization in NORT.** We systematically evaluated heuristic-based, configuration-specific deep learning, and combined hybrid approaches across

2-object and 5-object NORT settings. Training a unified behavior classifier on both configurations and subsequently resolving object identity through spatial reasoning produced the strongest overall generalization, demonstrating the benefit of separating shared visual behavior recognition from configuration-specific interpretation.

- **Characterization of transfer and adaptation across experimental settings.** We examined how components of the framework transfer when experimental conditions change. The Fisher 344 experiments evaluate adaptation to a substantial change in animal appearance, while the EPM and MBT extensions examine reuse of the framework under different arena geometries, objects, and behavioral objectives. Together, these experiments identify components that can be reused and components that require domain-specific fine-tuning or task-specific spatial reasoning.

- **End-to-end, analysis-ready behavioral processing at experimental scale.** We developed a complete computational pipeline that transforms raw behavioral videos into keypoint trajectories, frame-level behavior annotations, object- or region-specific measurements, temporal summaries, heatmaps, ethograms, and quantitative outputs for downstream neuroscience analysis. The NORT pipeline was applied to approximately 500 videos, demonstrating its practical use for large-scale behavioral analysis.

To accelerate further research on automated rodent behavior annotation, we will provide an open-source implementation of the proposed framework, including model weights and annotation tools for the NORT-focused pipeline. Furthermore, to promote reproducibility, we will make available a subset of the NORT videos used in this study, addressing a substantial need for publicly available datasets in automated rat behavior annotation. The EPM and marble-interaction extensions further illustrate how the same general workflow can be adapted to related behavioral paradigms with different spatial layouts, objects, and behavioral outcomes.

In addition to the open-source code, this work has resulted in an already published work and two to-be-submitted works.

- **Emily Alfs-Votipka**, Bhavana Sivayokan, Aliva Bakshi, Sanaz Gheibuni, Doina Caragea, Bethany Plakke, Dave Turner, and Daniel Andresen. "Automated behavior analysis in the novel object recognition test." Neurocomputing (2026): 134141.

- **Emily Alfs-Votipka**, Bhavana Sivayokan, Sanaz Gheibuni, Doina Caragea, Bethany Plakke, Dave Turner, and Daniel Andresen. "Toward Generalizable Vision-Based Rodent Behavior Analysis: Lessons and Guidelines from Cross-Domain Framework Extensions." In preparation. To be submitted to Scientific Reports.

- **Emily Alfs-Votipka**, Bhavana Sivayokan, Doina Caragea, Bethany Plakke, Dave Turner, and Daniel Andresen. "Automated Detection and Quantification of Marble Interaction in Rodent Behavioral Experiments." In preparation. To be submitted to Journal of Neuroscience Methods.

The primary computer science contribution of this dissertation is not a new pose-estimation or image-classification architecture in isolation, but a modular strategy for integrating learned visual perception with explicit, task-specific spatial reasoning. Existing systems emphasize complementary aspects of automated behavioral analysis. Markerless pose-estimation frameworks provide detailed localization but require a subsequent mechanism for assigning behavior; DeepEthogram (Bohnslav et al., 2021) learns behavioral labels directly from video; BehaviorDEPOT (Gabriel et al., 2022) emphasizes interpretable behavioral inference from pose trajectories; and EXPLORE demonstrates learned visual classification for object exploration. The framework developed here combines these complementary principles while explicitly separating components intended to capture transferable visual information from components that encode the geometry and behavioral definitions of a particular experiment.

Within this hybrid architecture, a compact representation consisting of the rat nose and tail-base provides the spatial information required for explicit behavioral reasoning. These keypoints provide information about animal position, body orientation, posture, and proximity to experimental objects. The resulting architecture allows learned models to perform visual recognition while explicit heuristics incorporate experimental geometry and domain knowledge that would otherwise need to be learned implicitly from image data.

A central methodological contribution is the separation of generic behavioral recognition from task-specific object identification. Rather than requiring a behavior classifier to learn a separate class for every object and arena configuration, the proposed NORT workflow first identifies generic `Object Interaction` behavior and subsequently uses keypoint-based spatial reasoning to determine the specific object involved. This factorization enables a single behavior-classification model to operate across both two-object and five-object configurations while retaining object-specific behavioral information. Experimental evaluation demonstrated that the combined model with heuristic refinement achieved the strongest overall generalization across these configurations.

A second research contribution is the empirical investigation of which components of the framework transfer across experimental conditions and which require domain-specific adaptation. Rather than assuming that a trained behavioral model transfers unchanged between datasets, this work evaluates adaptation to different object configurations, rat strains, arena geometries, and behavioral objectives. The Fisher 344 experiments demonstrate that visual domain changes can require targeted fine-tuning, while the EPM and MBT extensions demonstrate how the same computational components can be reused while replacing the task-specific spatial reasoning layer. This provides an empirical basis for understanding which components of an automated behavioral annotation pipeline can be transferred and which require domain-specific adaptation.

Finally, the dissertation contributes a complete computational pipeline that extends beyond model prediction to produce structured, analysis-ready behavioral data. The framework converts raw behavioral video into keypoint trajectories, frame-level behavioral labels, object interactions, temporal summaries, heatmaps, ethograms, and quantitative outputs suitable for downstream neuroscience analyses. The system was demonstrated at experimental scale through analysis of approximately 500 NORT videos. Thus, the contribution is not solely an individual machine-learning model, but a modular computational framework for transforming raw video into reproducible behavioral measurements.

Collectively, these contributions distinguish this work from prior systems by emphasizing the integration of learned perception, interpretable spatial reasoning, task-specific behavioral

inference, and cross-domain adaptation within a single workflow. The resulting framework provides a computational strategy in which general-purpose visual models can be combined with explicit representations of experimental structure, allowing the same underlying architecture to be adapted to different behavioral assays, animal appearances, arena geometries, and behavioral outcomes.

Chapter 2

Overview of YOLOv11

YOLOv11 represents a significant advancement in the YOLO (You Only Look Once) series, providing strong accuracy and real-time efficiency for modern vision workloads (Du et al., 2025; Gan et al., 2025; Luo and Matsumoto, 2025). While YOLO models were originally designed for object detection tasks (Redmon et al., 2016), more recent versions, including YOLOv11, are not limited to object detection, but offer competitive variants also for pose/keypoint estimation and image classification, all within a unified training and deployment stack (e.g., `yolo11n/s/m/l/x` with task-specific `-pose` and `-cls` heads). This breadth allows us to localize subjects when needed (e.g., rats), annotate fine-grained body landmarks (e.g., rat *nose* or *tail-base*), and classify frames (e.g., rat behavior classification) — all within the same framework. Architecturally, YOLOv11 continues the established three-stage design — *Backbone*, *Neck*, and *Task Head* (e.g., detection or classification head) - but refines each stage for speed and accuracy (Khanam and Hussain, 2024).

The *Backbone* serves as the primary feature extractor, systematically reducing the spatial resolution of the input image while increasing the channel depth to capture abstract and semantically rich features. YOLOv11 introduces the *C3k2* module in place of the earlier C2f block, using smaller kernel convolutions to achieve greater parameter efficiency and faster computation without sacrificing performance. The backbone also retains the *SPPF (Spatial Pyramid Pooling - Fast)* module from previous YOLO versions to aggregate multi-scale

contextual information, but it introduces a new *C2PSA (Cross-Stage Partial with Spatial Attention)* module. C2PSA enhances spatial attention, enabling the model to focus more effectively on important regions within the image and improving detection performance in complex visual scenes (Khanam and Hussain, 2024).

Features extracted by the *Backbone* are then passed to the *Neck*, which functions as a critical bridge for integrating multi-scale feature representations. The Neck applies upsampling and concatenation of feature maps from different levels, followed by additional *C3k2 blocks* to improve efficiency and enhance feature fusion. YOLOv11 further strengthens this stage with the inclusion of the *C2PSA attention mechanism*, allowing the network to emphasize salient spatial information (Khanam and Hussain, 2024).

The *Task Head* specializes predictions for bounding boxes and classes (detection), 2D keypoints (pose), or frame-level labels (classification), among others (Khanam and Hussain, 2024). The *Detection Head* processes the refined feature maps to produce bounding box coordinates, objectness scores, and class probabilities. The *Pose Head* extends detection to jointly regress per-instance keypoint coordinates, their visibility and confidence scores, for each detected bounding box. The *Classification Head* aggregates global features (e.g., global average pooling) and outputs frame-level class logits and calibrated probabilities via a lightweight linear classifier. The general YOLOv11 architecture and the task heads employed in our study are depicted in Figure 2.1.

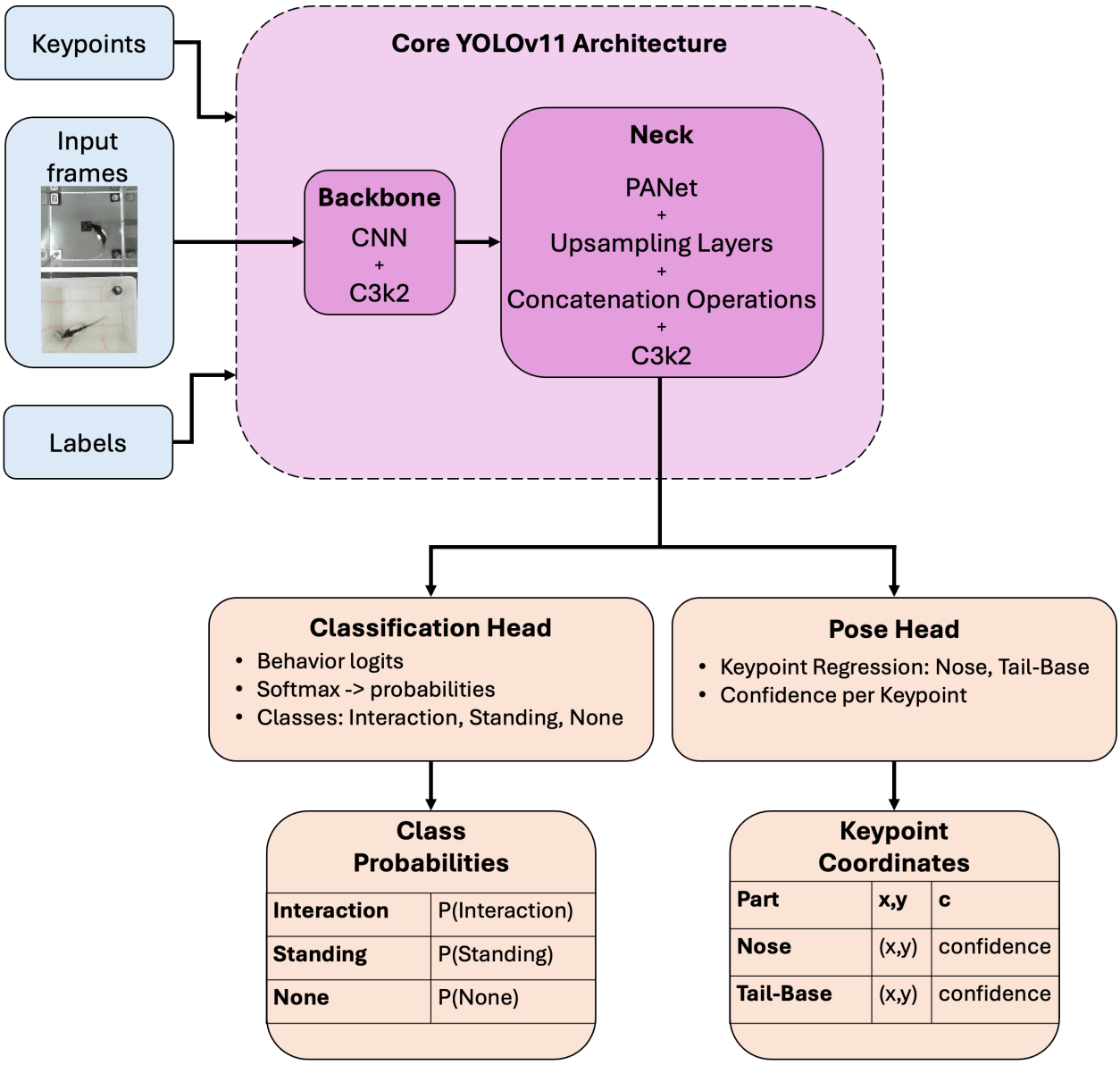


Figure 2.1: *General YOLOv11 architecture and different task heads employed in this study.*

Chapter 3

Video data collection

The behavioral neuroscience datasets used in this dissertation were generated by Dr. Bethany Plakke's lab at Kansas State University. The primary dataset consists of NORT videos collected from Long-Evans rats used in a model of fragile X syndrome, with additional videos collected for EPM and MBT analyses. Long-Evans rats are a commonly used laboratory strain characterized by a distinctive pattern consisting of a predominantly white coat with a dark, hood-like marking that extends from the head along the dorsal surface of the body. A typical individual can be seen in Figure 3.1.

In addition to the primary NORT dataset, a separate set of NORT videos was collected using a different rat strain. These animals were Fisher 344 rats, which are characterized by a uniform white coat. A representative individual is shown in Figure 3.2.

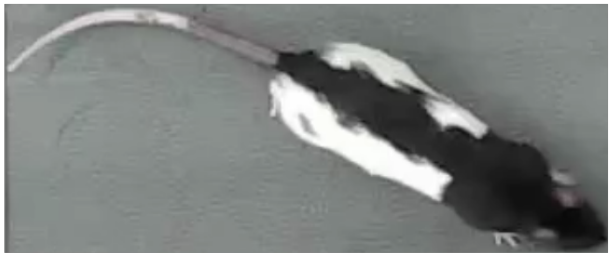


Figure 3.1: *Long-Evans rat. Individual Long-Evans rat with the breed-characteristic white body and black hood down the spine of the rat.*



Figure 3.2: *Fisher 344 rat. Individual Fisher 344 rat with the uniform white coloration.*

3.1 NORT videos

The NORT test is used to explore the cognitive deficits of knockout rats with the genetic abnormality causing FXS by assessing their memory in comparison to wildtype rats. Data obtained from these videos will allow the behavioral neuroscientists to compare the amount of time the knockout and wildtype rats spent with the novel objects over the familiar objects.

Over 500 trial videos were recorded for both 2-object and 5-object NORT tests, using a top-view recording format as shown in Figure 1.1. Each video in the dataset is approximately 10 minutes long with a resolution of (1920×1080) pixels and a frame rate of 60 frames per second (fps). In the 2-object videos, there is a Long-Evans rat in a rectangular enclosure, while in the 5-object videos, the Long-Evans rat is in a square enclosure. Given the resolution and known distances in the videos, these trials have pixel-per-centimeter (ppcm) values of 13.95ppcm and 10.45ppcm for the two and five object trials respectively.

The objects used had different visual features and were made of either metal or glass, to ensure that they were not chewable and did not retain any scent cues. The objects were fixed on metal platforms, which were then fixed to the floor of the cage using Velcro. In the rare case that a rat was able to dislodge an object, the trial was terminated and the video was eliminated from the dataset. The arena and objects were cleaned between each animal using Peroxigard to sanitize and remove scent cues.

The 2-object NORT test was performed over a period of three days, and it was done for two timepoints, when the rats were young (postnatal day 50-65) and also when the rats were old (12 months). The procedure was similar at both timepoints. The sequence of days is illustrated in Figure 3.3 (Top Row). On the first day, the rats were allowed to explore an empty arena for 10 minutes. On days two and three, two objects were placed in the arena

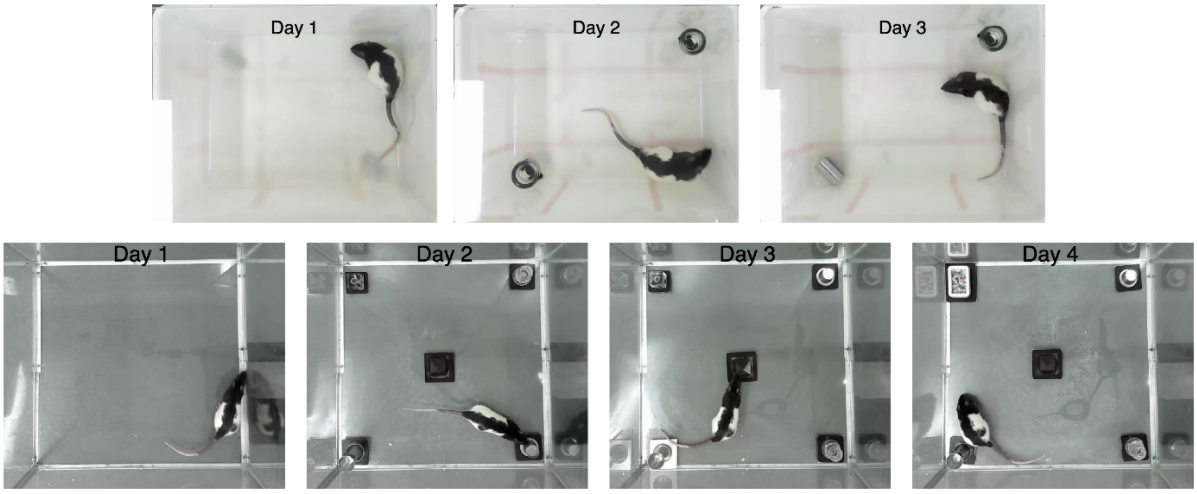


Figure 3.3: *NORT setup for 2-object and 5-object videos. (Top Row) Frames illustrating the setup across the three days of the 2-object NORT test. On the first day, the rat explores the empty enclosure. Two objects are introduced on the second day. On the third day, a new object replaces one of the original objects. (Bottom Row) Frames illustrating the setup/changes across the four days of the 5-object NORT test. On the first day, the rat explores the empty enclosure. Five objects are introduced on the second day. The two objects on the right swap positions on the third day. Finally, a new object replaces the original object in the upper left corner on the fourth day.*

and the animals were allowed to explore the arena and objects for 10 minutes. Specifically, on the second day, two identical objects were introduced in the arena. For the young age group, these objects could either be metal sockets or glass candle holders. For the old age group, these objects could be ramekins or glass dice. On the third (final) day, the arena had an object that the rats saw the previous day and a novel object. In addition to the primary Long-Evans NORT dataset, a smaller set of Fisher 344 NORT videos was collected later in the project using a similar 2-object NORT enclosure and protocol. These videos were used as an extension dataset to evaluate whether the Long-Evans-trained keypoint and behavior models could be adapted to a visually distinct rat strain.

The 5-object NORT test was carried out over a period of four consecutive days as illustrated in Figure 3.3 (Bottom Row) when the rats were young (postnatal day 60-75). On the first day, the rats were allowed to habituate to an empty arena for 10 minutes. On days two through four, five objects were placed in the arena and the animals were allowed to explore

the arena and objects for 10 minutes. On the second day of the test, one object was placed at each of the four corners of the arena, and the fifth object was placed at the center of the arena. On the third day, two objects were swapped places while the other three objects were kept the same. On the final day of the test, the arrangement of objects was the same as that on the third day, but one of the objects was replaced by a novel object. The objects and their positions were not counterbalanced in the 5-object videos, due to the inherent complexity introduced by a larger number of objects.

3.2 EPM videos

The elevated plus maze (EPM) is widely used to assess anxiety in rodents and has been shown to have face, construct, and predictive validity in measuring anxiety in rodents. This four-armed maze, approximately 50cm off the ground, has two open arms and two closed arms arranged in the form of a cross with similar arms opposite each other and makes use of rodents' natural aversion to open and elevated areas to assess anxiety. A related conflict-test variant was also conducted using the same EPM arena, but with the open arms baited with food. This variant creates a conflict between avoiding the exposed open arms and approaching a food reward. In both these tests, the time spent in the closed arms over open arms is indicative of the anxiety level of the animal, with more time in the closed arms than open arms indicating higher levels of anxiety. The EPM and conflict tests were used to assess the level of anxiety in the knockout rats having the genetic abnormality causing FXS. To begin the EPM or conflict test, the rats were placed in the center of the maze facing an open arm away from the researcher. They were then allowed to explore the maze for 5 minutes and the trial was videotaped at a resolution of (1920×1080) pixels and a frame rate of 60 frames per second (fps) using a top view recording format. Given the resolution and known distances, this trial has 8.381ppcm. The maze was cleaned with Oxivir between animals to sanitize and remove scent cues. During the conflict test, which was carried out on two consecutive days following the EPM test, the open arms were baited with a few pieces of Honey Nut Cheerios™ (General Mills).

3.3 MBT videos

The marble burying test (MBT) is used to assess repetitive behavior in rodents and takes advantage of the rodents' natural tendency to dig in natural settings. Repetitive behavior is assessed from the amount of time spent interacting with and burying marbles. This test was carried out in a rectangular arena containing regular aspen chip bedding. The rats were habituated to this arena for 5 minutes on the day before the test as well as immediately before the test. After habituation on the day of the test, the rats were briefly transferred to their home cage, and 20 marbles were placed in a 5 x 4 grid in the testing arena. The rats were then placed back in the arena and given 10 minutes to explore and bury the marbles. This trial was videotaped at a resolution of (480×480) pixels and a frame rate of 15 frames per second (fps) using a top view recording format. Given the resolution and known distances, this trial has 6.32ppcm.

At the end of 10 minutes, the number of marbles buried at least 2/3 into bedding were counted. Once animals were returned to their home cage, the marbles were retrieved from the bedding and cleaned with Oxivir to sanitize and remove scent cues. The bedding was dumped and the arena was also cleaned with Oxivir. Although the standard endpoint measure is the number of buried marbles, the video recordings also make it possible to analyze the timing and sequence of rat-marble interactions throughout the trial, which motivates the marble-interaction analysis described in Chapter 10.

Chapter 4

Keypoint annotation

Each of the experimental paradigms examined in this work relies upon accurate keypoint annotation as a foundational component of the behavioral analysis pipeline. Accordingly, this chapter presents the development, training, and evaluation of an initial model designed to annotate the *nose* and *tail-base* keypoints of Long-Evans rats. These keypoint predictions serve as the basis for the spatial analyses described throughout this dissertation as well as a baseline comparison for the behavior classification.

As the project progressed, additional behavioral datasets and experimental paradigms were introduced in response to evolving research objectives and collaborator requirements. While the initial model was developed using a core dataset, subsequent efforts focused on adapting and extending the framework to accommodate these new experimental conditions. The methodologies and modifications required to support these additional datasets are discussed in Chapter 9.

4.1 YOLOv11 pose model for rat keypoint annotation

For the task of keypoint annotation, the YOLOv11 Pose model with the medium (`m`) configuration (`yolo11m`) and its default settings and hyperparameters was employed. The default settings are commonly used as starting points when fine-tuning YOLOv11 Pose (Ultralyt-

ics YOLO, 2025). The training process was conducted over a maximum of 2000 epochs. However, to prevent overfitting and to enhance generalization performance, early stopping was implemented with a patience threshold of 100 epochs. A mini-batch size of 16 was selected, balancing the trade-off between computational efficiency and model convergence stability. Each input image was resized to a fixed resolution of (640×640) pixels and mosaic augmentation was used.

By default, the YOLOv11 Pose model monitors five distinct loss components during training to evaluate and optimize model performance. Each loss contributes to the overall objective function, weighted according to its relative importance in the learning process. The default losses and their respective weights were used. Specifically, the training objective of the YOLOv11 pose model is defined as:

$$L_{\text{total}} = \lambda_{\text{box}} L_{\text{box}} + \lambda_{\text{pose}} L_{\text{pose}} + \lambda_{\text{obj}} L_{\text{obj}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dfl}} L_{\text{DFL}} \tag{4.1}$$

Bounding box regression loss L_{box} was calculated using the Intersection over Union (IoU) metric and assigned a weight λ_{box} of 7.5, reflecting its significant influence on accurate localization of objects. Pose loss L_{pose}, responsible for supervising keypoint localization, was heavily emphasized with a weight λ_{pose} of 12.0, highlighting the centrality of pose estimation to the model's task. Objectness loss L_{obj}, used to determine whether an object is present in a given region, was monitored with a weight λ_{obj} of 1.0. Classification loss L_{cls} was computed using Binary Cross-Entropy (BCE) with a weight λ_{cls} of 0.5, indicating a relatively lower emphasis on class prediction accuracy compared to localization. Distribution Focal Loss (DFL) L_{DFL}, which aids in more precise bounding box regression by modeling the distribution of target locations, was weighted at λ_{DFL}=1.5. These weighted components collectively guide the training process, enabling the model to effectively balance object detection and keypoint estimation.

After the completion of model training, it was observed that a subset of frames still exhibited missing keypoint annotations. Missing labels typically occurred when the nose was occluded by an object or when the subject was in a standing posture, thereby obscuring

the tail-base. Examples of particularly difficult frames can be seen in Figure 4.2. Post-processing was applied to video frames lacking annotations for the *nose* and/or *tail-base* of the rat. To address such discontinuities, missing nose or tail-base coordinates were estimated using linear interpolation between the nearest preceding and succeeding frames with valid keypoint predictions. For gaps occurring at the beginning or end of a video, the nearest available valid prediction was used.

4.2 Manual frame-based keypoint annotation

Manual annotation for training and evaluating rat keypoint/pose models was performed in LabelMe (Russell et al., 2008). Each annotated frame includes two keypoints - *nose* and *tail-base* - and a bounding box around the entire rat; a screenshot of the process is shown in Figure 4.1.



Figure 4.1: *LableMe interface. Screenshot of the interface for keypoint annotation using LabelMe.*

The *nose* and *tail-base* of the rat can typically be readily identified. However, there are cases when the *nose* and/or *tail-base* are obscured by objects in the arena or due to the nature of pose (e.g., grooming, huddled). When landmarks were occluded, we used the following fallback rules: for *nose* - annotate the most anterior visible point on the snout; if the rostrum is fully hidden, advance posteriorly to the interaural line as a *nose* surrogate; for *tail-base* - annotate the most proximal visible point along the midline (spine or proximal tail) aligned with the animal's orientation.

Examples of complex annotations can be seen in Figure 4.2.

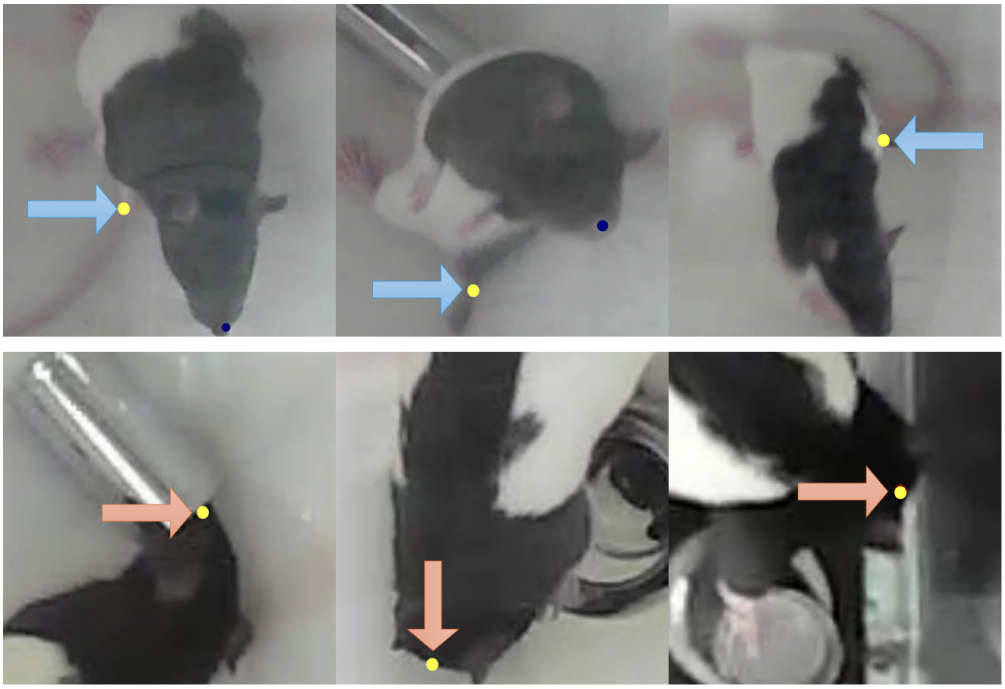


Figure 4.2: *Challenging poses for keypoint annotation. In some cases, annotation of the nose and tail-base can be complicated by the posture of the rat. The top row with blue arrows shows difficult frames for tail annotations and where the annotation was placed. As a guideline, the point nearest the tail-base, either further down the tail or up the spine, was selected. The bottom row with orange arrows shows difficult frames for nose annotations and where the annotation was placed. When the nose was obscured, the point nearest the nose going up the snout was selected. In some cases, this went as far as between the ears.*

We built the keypoint dataset via an iterative bootstrapping loop: starting with ∼150 manually annotated frames, we trained a model, ran inference on unlabeled frames, and then visually inspected predictions to identify failure cases for additional manual annotation. To

ensure diversity and generalizability, frames were drawn from eleven videos: four 2-object NORT videos (one from day 1, one from day 2, and two from day 3), six 5-object NORT videos (one from day 1, one from day 2, two from day 3, and two from day 4), and one non-NORT video (specifically, an EPM video) featuring the same Long–Evans rat strain. This loop—training, inference, visual evaluation, and targeted frame selection—was repeated four times with a consistent training procedure, yielding a final dataset of 2,550 annotated frames: 2,200 frames from NORT videos, equally split between 2-object and 5-object videos, and 350 frames from the EPM video. Frames for the test set were selected from held-out videos not used for training or validation. These frames were first annotated by a preliminary model and then manually corrected. The final distribution of the dataset is summarized in Table 4.1.

Video Type	#Objects	Video	Type/Day	Train	Val	Test	Total
NORT	2-object	F1a.14_G1	young, day 1	132	34	0	166
		F8a.03_G2	young, day 2	133	34	0	167
		F5a.10_G3	young, day 3	134	33	0	167
		B6.03_G3	young, day 3	0	0	600	600
		Total 2-object		399	101	600	1,100
	5-object	F1a.04_B1	young, day 1	29	7	0	36
		F8a.03_B2	young, day 2	29	7	0	36
		F5a.10_B3	young, day 3	29	7	0	36
		F5a.03_B4	young, day 4	241	60	0	301
		F5a.10_B4	young, day 4	73	18	0	91
		B6.04_B3	young, day 3	0	0	600	600
		Total 5-object		401	99	600	1,100
Non-NORT		F1.02_epm	old	40	10	300	350
		Total combined		840	210	1,200	2,550

Table 4.1: *Keypoint annotated dataset. Information about the NORT (2-object and 5-object) videos and Non-NORT (EPM) video used to select frames, as well as the type or rat (young or old) in these videos is provided. Statistics about the frame composition of the training (train), validation (val) and test subsets are also shown, independently for each video, for each video type and also overall for the whole combined set of frames.*

4.3 Metrics for keypoint annotation

We evaluated the YOLOv11 Pose model using the standard metrics reported by the Ultralytics pose-estimation framework, including mAP@0.50 and mAP@[0.50:0.95]. These met-

rics summarize localization performance across confidence and localization thresholds. For mAP@0.50 and mAP@[0.50:0.95], values closer to 1 indicate higher accuracy. Because the downstream behavioral heuristics depend directly on the accuracy of the predicted nose and tail-base coordinates, we also report coordinate-level error analyses in millimeters for the manually corrected test frames.

4.4 YOLOv11 keypoint annotation model performance

The model achieved an mAP@0.50 value of 87.4% and an mAP@[0.50:0.95] value of 87.3%. While the keypoint detection model demonstrated strong overall performance in terms of mAP, some instances of missed keypoints were observed. These omissions occurred either in isolation, where a single keypoint was not detected, or concurrently, with both *nose* and *tail-base* keypoints missing within the same frame. Upon reviewing frames with missing annotations, it was noted that these errors generally corresponded to instances in which the keypoints were visually obscured or when the rat appeared slightly out of focus, thereby reducing visibility. Table 4.2 shows the distributions of missing keypoints over the four test videos from our NORT behavior dataset.

In both of the 2-object videos, only one or the other keypoint was ever missing. However, in the 5-object videos, there were some instances of both keypoints missing in the same frame as well as single keypoints missing. Of the two keypoints, the *nose* was missing more often than the *tail-base*. In addition to evaluating keypoint localization accuracy on the manually corrected test frames summarized in Table 4.1, we also assessed missing keypoint predictions over the full-length NORT behavior-classification test videos used in Chapter 5.

However, across all four videos, with over $144,000$ frames, only 280 frames had any missing annotations, which represents a very small percentage of frames out of the total number of frames in the test videos used.

We further analyzed the missing frames by identifying and quantifying continuous sequences of missing annotations. Figure 4.3 illustrates the frequency and length of these gaps across the behavior annotation test dataset. Across the four test videos, the longest contin-

#Objects	Video	#Frames	Nose	Tail	Both	Total
2-object	B06_G3	36,000	34	2	0	36
2-object	F8.16_G2.3	36,024	13	0	0	13
5-object	B10_B2	35,999	21	1	73	95
5-object	F1a.09_B3	35,988	17	17	102	136
	Total	144,011	85	20	175	280

Table 4.2: *Missing keypoint annotations. Over the test dataset for behavior classification, we analyzed the number and type of missing keypoints. Each column indicates how many frames were missing annotations of that type. The* ***Both*** *column displays how many frames were missing both nose and tail-base keypoints.*

uous frame sequence with missing *nose* and/or *tail* annotations consisted of 43 frames and was found in a 5-object video. Given the recording frame rate of 60 fps, this corresponds to approximately 7/10 of a second of uninterrupted data loss. The next longest sequence had 24 frames missing annotations also in a 5-object video. For the two object videos, 10 was the longest number of consecutive frames with missing annotations.

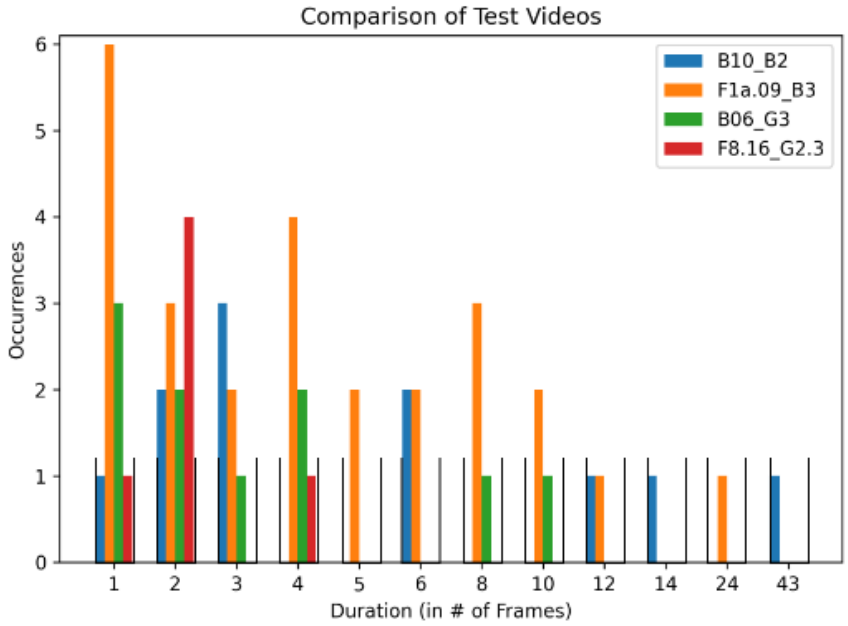


Figure 4.3: *Duration of sequential missing keypoints. For the test dataset of behavior classification videos, we examined the number of sequential missing frames. As an example, video F1a.09_B3 in orange had two instances of 6 sequential frames missing. With a frame rate 60fps, this is a tenth of a second.*

To mitigate the effect of missing data on behavioral metrics, an interpolation strat-

egy was employed for both *nose* and *tail-base*. We analyze the results for *nose* annotation/interpolation in what follows. Similar results were also observed for *tail-base* annotation/interpolation. Specifically, for frames in which the nose position was undetected, its location was estimated using linear interpolation from temporally adjacent frames - both preceding and succeeding the gap. This approach assumes smooth motion continuity and helps preserve the temporal coherence of the animal's trajectory. While this method introduces some degree of approximation, it was deemed acceptable for short gaps based on visual inspection and preliminary quantitative validation. The results of the final model with interpolation for nose keypoint can be seen in Figure 4.4. In the 2-object video, spanning 600 frames, the maximum distance between predicted and ground-truth nose annotations was 133 mm, with 90% of predictions within 20 mm. For the 5-object video, also comprising 600 frames, the maximum deviation was 80 mm, and 90% of predictions were within 14 mm of the ground truth.

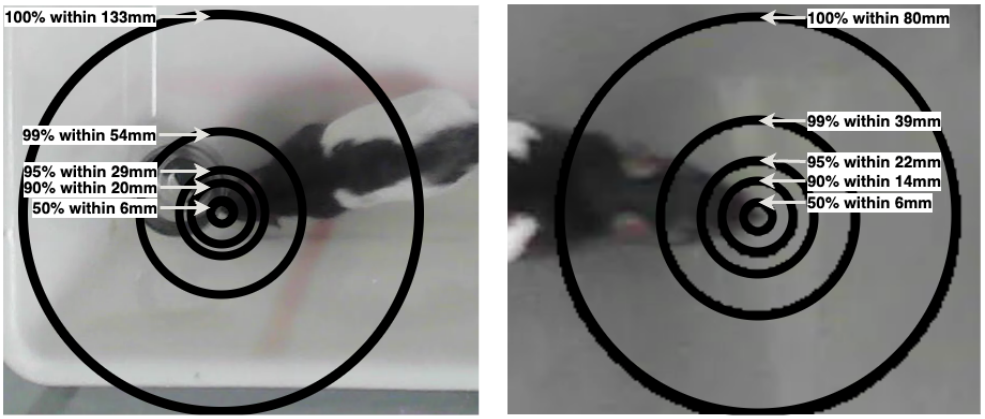


Figure 4.4: *Nose annotation analysis post-interpolation. The graphics show the distribution of distances from the ground truth nose (center) to predicted nose after interpolation. For the 2-object video (left), we can see that 99% of the nose predictions are within 54 mm of the ground truth over 600 frames. For the 2-object test set, the largest distance was 133 mm. For the 5-object video (right), we see that 99% of the nose predictions are within 39 mm of the ground truth and the largest distance was 80 mm over 600 frames.*

A few samples of the resulting interpolation of the *nose* can be seen in Figure 4.5. Good *nose* interpolation results are seen when only a small number of frames in a sequence are missing annotations (see examples in the left column of Figure 4.5). As the number of

sequential frames missing annotations increases, the quality of the *nose* interpolation degrades (see examples in the right column of Figure 4.5). In the videos from which these examples were selected, the rat is typically in the process of standing up or returning to all four paws. With this rapid change of motion, the number of missing annotations increases and the interpolation accuracy decreases. Despite some keypoint errors, the overall results of our keypoint annotation model are sufficient for the subsequent heuristic-based behavior annotation tasks.

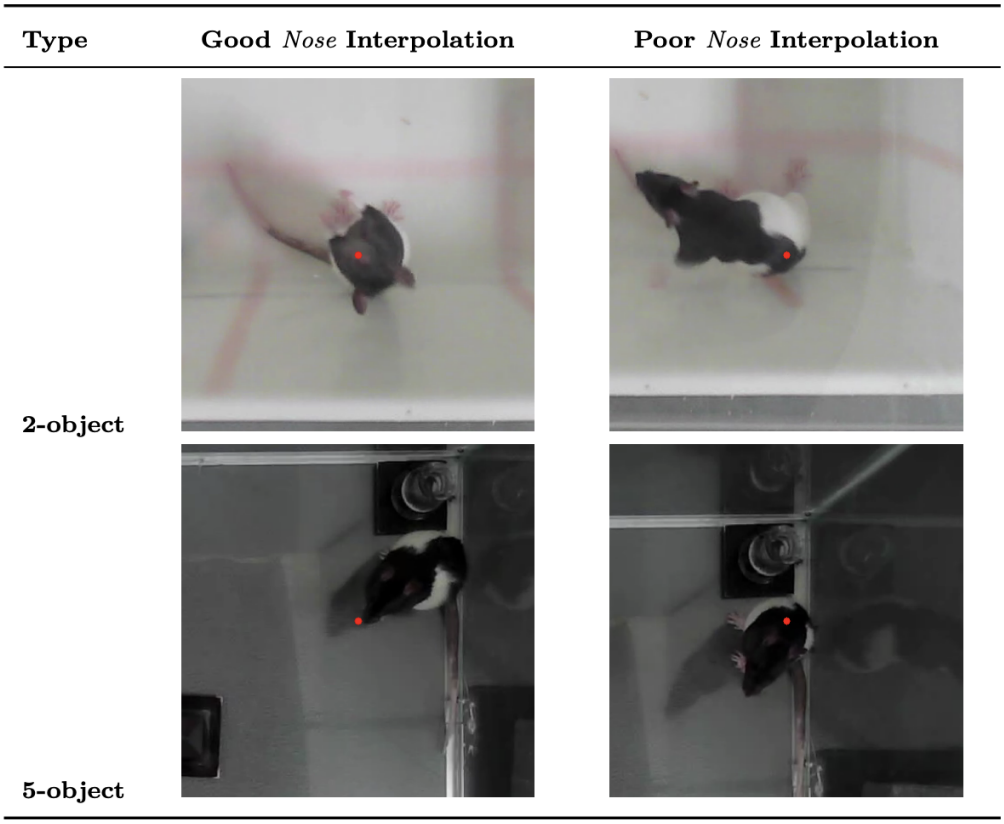


Figure 4.5: *Sample interpolation results for the nose keypoint. Frames with missing nose keypoint annotations are further processed by estimating the nose position using linear interpolation between the nearest preceding and succeeding frames with valid nose predictions. The left column illustrates good interpolation results, while the right column contains poorer interpolation results.*

Overall, the YOLOv11 Pose model provided reliable *nose* and *tail-base* localization across the NORT test videos, with missing predictions occurring in only a small fraction of frames and primarily under occlusion or difficult postural conditions. These results support the

use of the predicted keypoints as inputs to the spatial heuristics and behavior-classification refinements described in Chapter 5.

Chapter 5

Novel object recognition test (NORT)

5.1 NORT behaviors of interest

The time each rat spent with each of the objects needs to be tracked to determine whether the rats were able to detect the changes in the objects and their locations. If the rats had intact memory for the objects and their locations on the previous days, they were expected to spend more time with the objects that had swapped places (on the third day in the 5-object videos) and the new object on the final day of either the 2-object or the 5-object videos. To be able to track the time each rat spent with each of the objects, each frame needs to be annotated according to specific object interactions, i.e., `Upper` or `Lower` for 2-object videos, and `Object 1` to `Object 5` for 5-object videos, globally referred to as `Object Interaction` behaviors. Furthermore, the neuroscientists were also interested in the `Standing`, `Resting`, and `Freezing` behaviors, while any other behaviors are referred to as `Other`.

A heuristic approach based on *nose/tail-base* keypoints was originally used to identify the behaviors. For `Resting` and `Freezing` behaviors, established heuristics based on a sequence of frames were used and visual examination of the results by domain experts showed satisfactory results. For the `Object Interaction` and `Standing` behaviors, frame-level keypoint-based heuristics were designed. However, preliminary examination by domain experts suggested that these behaviors needed further improvement. Thus, the `Object Interaction`

and `Standing` behaviors versus all the `Other` behaviors are the focus of the models developed in this work. In what follows, we describe the manually annotated behavior dataset used to train and evaluate the YOLOv11 frame-level behavior classification models. We then describe the keypoint-based heuristics, the YOLOv11 classification models, and the evaluation protocol used to compare heuristic-based and model-based NORT behavior annotation.

5.2 Manual frame-level behavior annotation

Manual frame annotation for training and evaluating behavior recognition models was judiciously performed by a domain expert using the Computer Vision Annotation Tool (CVAT) (CVAT.ai Corporation, 2023) available at `https://github.com/cvat-ai/cvat`. To simplify the annotation process, we used the fact that a behavior of interest is typically sustained for a number of frames and only annotated the beginning of a behavior and the end of that behavior with tags such as *begin-behavior*/*end-behavior*; a screenshot of the process is shown in Figure 5.1.

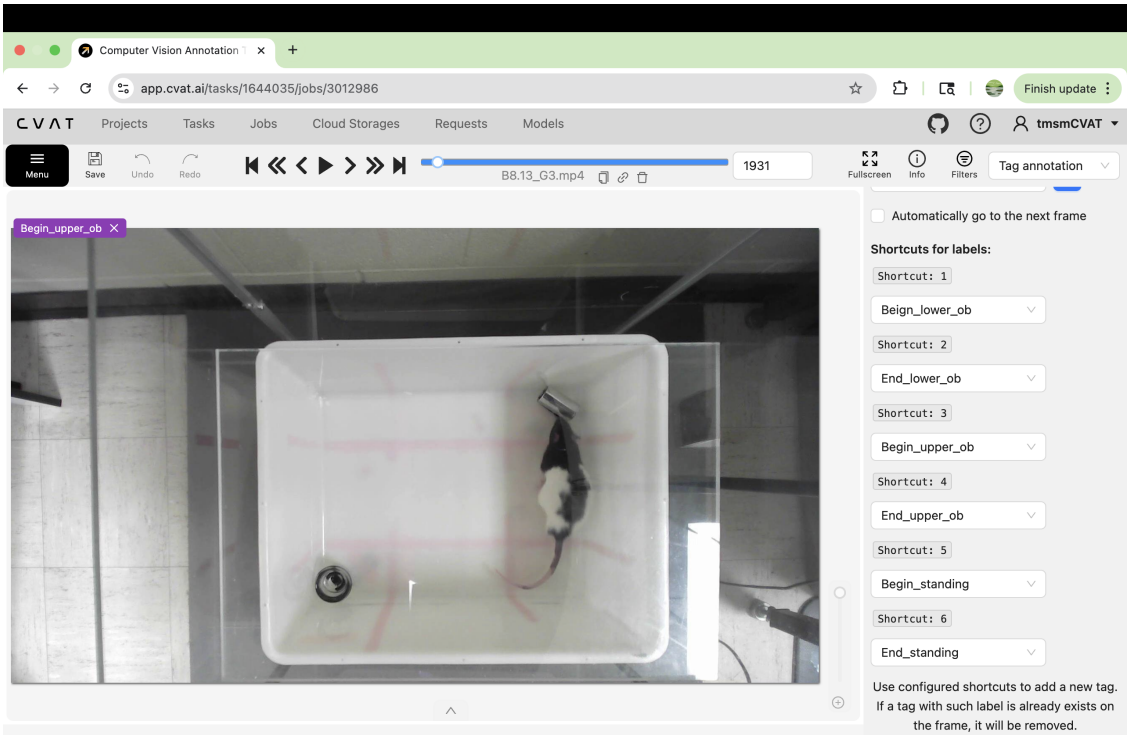
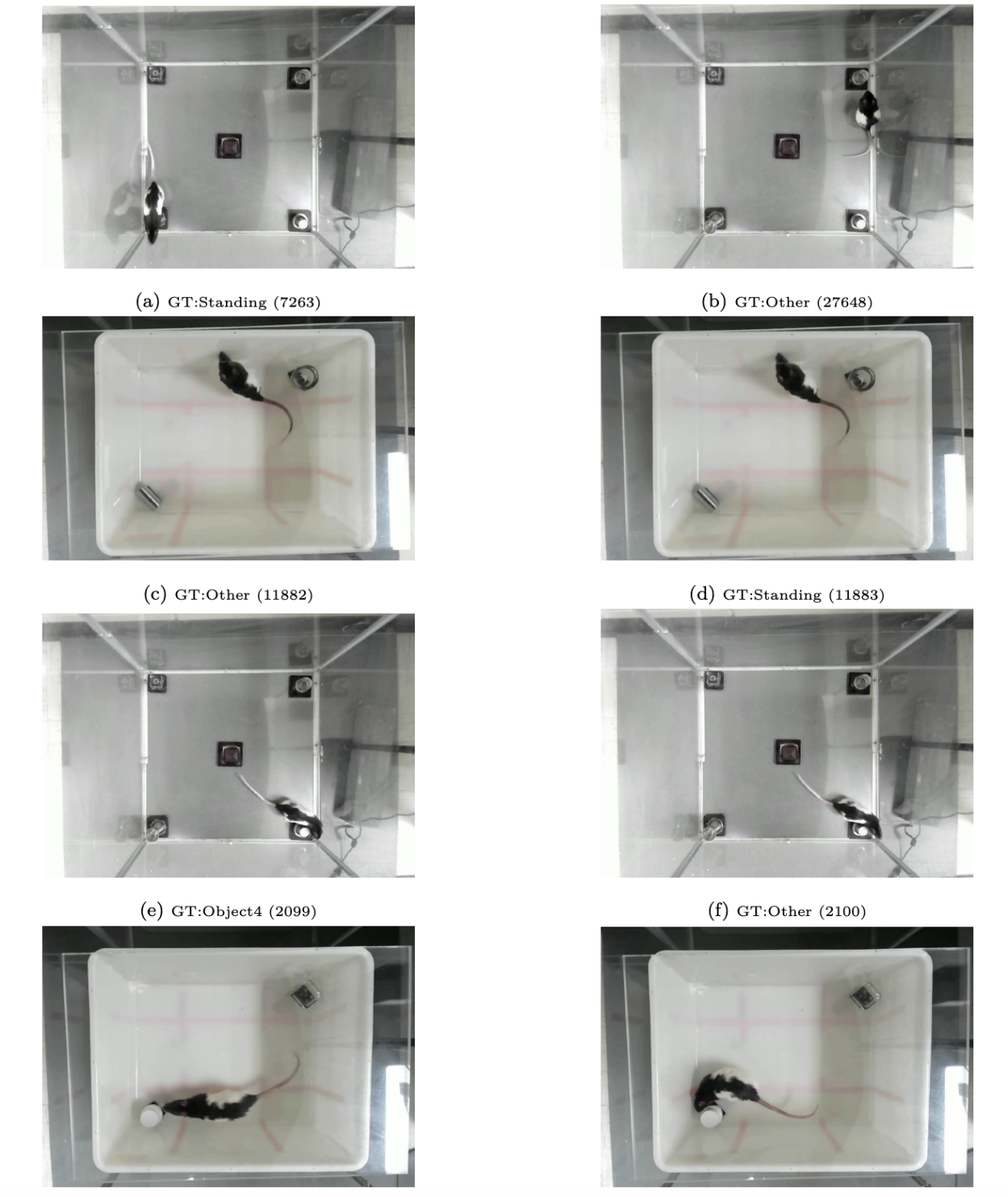


Figure 5.1: *Screenshot of the interface for behavior annotation using CVAT.*

Subsequently, we used a custom script to transform the CVAT output file into a file that can be used to train the YOLOv11 models. Specifically, we used the script to label all the frames in between *begin-behavior*/*end-behavior* tags with the respective behavior, while all the other frames were annotated as `Other`.

The following guidelines were used for annotation. If the rat was approximately within 1 cm of the object and was oriented towards the object, the behavior was marked as `Object Interaction`. Stepping on the object while standing up on hind legs, playing with the velcro that was used to attach the object to the floor of the arena, or pulling on the metal platform to which objects were attached were not counted as object interaction. Despite proximity to the object, if the animal did not look at, sniff, or lick the object, it was not considered as object interaction. `Standing` (a.k.a., `Rearing`) was annotated when rats were on their hind legs. This predominantly included instances where the rats reared close to the walls of the arena with the forelimbs resting on the walls or object, but also included instance of rearing in the middle of the arena where the forelimbs were unsupported. Transition from four legs to two legs was marked as the beginning of `Standing` behavior, while transition to all fours was marked as the ending of `Standing` behavior. Only one behavior was annotated for any given frame. Some examples illustrating challenges of behavior annotation are shown in Figure 5.2.



(a) GT:Standing (7263)

(b) GT:Other (27648)

(c) GT:Other (11882)

(d) GT:Standing (11883)

(e) GT:Object4 (2099)

(f) GT:Other (2100)

(g) GT:Other (6989)

(h) GT:Lower (7194)

Figure 5.2: *Hard annotations at the boundary between behaviors. Each subpanel shows a single frame with the ground-truth (GT), along with the frame index. Frames in a row are from the same video.*

We should note that the `Freezing` and `Resting` behaviors were not manually annotated, as they needed to be identified based on duration-dependent (specifically, at least one second or equivalently 60 frames) immobility rules and it was challenging to manually count at least 60 frames. At the same time, duration-dependent immobility rules based on keypoints were easy to code and provided accurate classification of `Freezing` and `Resting` behaviors based on domain expert review of the annotations. Therefore, we used this approach to annotate the `Freezing` and `Resting` behaviors.

A small subset of videos was selected for manual annotation. The selection of the videos was guided by two main goals: 1) to ensure a large and diverse training subset that includes both 2-object and 5-object NORT videos, as well as young and old rats; 2) to ensure that most settings are covered in the test set used to evaluate the performance. Specifically, we annotated 3 videos for training, 3 videos for validation and 4 videos for test.

The characteristics of the videos selected for annotation, as well as precise statistics about the frames annotated in each category and each video, are shown in Table 5.1, which also shows statistics for each subset and globally for the whole dataset. Overall, ∼360k frames were annotated and included in the dataset, with a class distribution of 1:1:2 for the `Object Interaction`, `Standing` and `Other` categories, respectively. Table 5.2 and Table 5.3 show statistics for the specific object interactions for the 2-object and 5-object videos, respectively.

Subset	#Objects	Video	Type/Day	Obj. Int.	Standing	Other	Total
Train	2-object	F1a.02_G3	young, day 3	13,459	10,165	12,364	35,988
	2-object	F8.03_G2.3	old, day 3	6,617	7,950	21,442	36,009
	5-object	F5a.03_B4	young, day 4	9,004	11,634	15,381	36,019
		Total		29,080	29,749	49,187	108,016
Val	2-object	B8.13_G3	young, day 3	6,497	7,405	22,098	36,000
	2-object	B8.02_G2	young, day 2	8,515	15,006	12,479	36,000
	5-object	B6.04_B3	young, day 3	16,428	5,092	14,480	36,000
		Total		31,440	27,503	49,057	108,000
Test	2-object	B06_G3	young, day 3	7,924	14,220	13,881	36,025
	2-object	F8.16_G2.3	old, day 3	2,801	8,413	24,811	36,025
	5-object	B10_B2	young, day 2	10,535	6,870	18,595	36,000
	5-object	F1a.09_B3	young, day 3	8,398	11,914	15,669	35,981
		Total		29,658	41,417	72,956	144,031
All				90,178	98,669	171,200	360,047

Table 5.1: *Behavior annotated dataset. Videos selected for each of the training (Train), validation (Val), and Test subsets, together with their characteristics and statistics about the number of frames annotated in each video and each category (`Object Interaction` - Obj. Int., `Standing` and `Other`), as well as for each subset and globally for the dataset.*

Subset	Video	Upper Obj	Lower Obj	Total
Train	F1a.02_G3	1,885	11,574	13,459
	F8.03_G2.3	4,583	2,034	6,617
	Total	6,468	13,608	20,076
Val	B8.13_G3	2,420	4,077	6,497
	B8.02_G2	3,832	4,692	8,524
	Total	6,252	8,769	15,021
Test	B06_G3	5,511	2,413	7,924
	F8.16_G2.3	1,231	1,570	2,801
	Total	6,742	3,983	10,725
All		19,462	26,360	45,822

Table 5.2: *Annotated behavior dataset (`Upper Object`/`Lower Object`). Statistics about specific object interactions (`Upper Object` - Upper Obj, and `Lower Object` - Lower Obj.) in the training (Train), validation (Val), and Test subsets, respectively, for each of the 2-object videos independently and also globally for the dataset (All).*

Subset	Video	Object1	Object2	Object3	Object4	Object5	Total
Train	F5a.03_B4	1,557	1,317	2,960	1,376	1,794	9,004
Val	B6.04_B3	3,943	2,052	3,877	3,988	2,568	16,428
Test	B10_B2	891	2,215	2,854	1,801	2,774	10,535
	F1a.09_B3	1,546	1,782	1,501	1,571	1,998	8,398
	Total	2,437	3,997	4,355	3,372	4,772	18,933
All		7,937	7,366	11,192	8,736	9,134	44,365

Table 5.3: *Annotated dataset (`Object 1` to `Object 5`). Statistics for specific object interactions (`Object 1` - Obj1 to `Object 5` - Obj5) in the training (Train), validation (Val), and Test subsets, respectively, for each of the 5-object videos independently and also globally for the dataset (All).*

To assess the difficulty of the annotation task and establish human error estimates for the NORT behavior classification task, two of the validation videos B8.13_G3 and B6.04_B3 were also annotated by a student, in addition to the domain expert. The student was given the guidelines designed by the domain expert and was allowed to annotate several minutes of a practice video to familiarize themselves with the guidelines and the annotation tool (CVAT). After the practice annotation was complete, the student's work was checked to ensure that the guidelines were properly followed and that there were no overlapping behaviors. Feedback was provided, as needed. Subsequently, the student completed the annotations of two validation videos, a 2-object video and 5-object video.

To quantitatively assess the level of agreement between annotators, the Cohen's kappa

coefficient was calculated for the annotated videos. Cohen's kappa provides a measure of inter-rater agreement between two annotators while accounting for the level of agreement that could be expected to occur by chance. In this analysis, agreement between the domain expert annotations and the student-generated annotations was evaluated on a frame-by-frame basis. A Cohen's kappa coefficient of 1.0 represents perfect agreement between the two annotators across all evaluated frames, whereas a coefficient of 0 indicates agreement equivalent to that expected by chance. The comparison between the domain expert and student annotations yielded a Cohen's kappa coefficient of 0.872 for the two-object video. For the five-object video, the coefficient was slightly lower at 0.863.

5.3 Keypoint-based heuristics for frame-level behavior classification

Keypoint-based heuristics were manually designed with domain expert input to label the behaviors. The behavioral categories identified through the keypoint-based heuristics, listed in order of annotation priority, include: `Standing`, `Object Interaction`, `Resting`, and `Freezing`. Additionally, for `Object Interaction`, the object the rat was interacting with was also identified.

Two distinct but complementary rule-based criteria were applied to identify the `Standing` behavior: (1) wall-supported standing, and (2) unsupported or free-standing within the center of the enclosure. For the first case - wall-supported standing - only the *nose* position was required. This classification additionally relied on a predefined bounding box that delineates the outer boundary of the arena. Specifically, the first frame of the video was extracted, and the boundaries of the enclosure were manually annotated using the LabelMe tool (Russell et al., 2008). If, in a given frame, the nose position was detected outside of the enclosure's bounding box, the animal's behavior was classified as `Standing`. The second type - free-standing - required both *nose* and *tail-base* positions to evaluate body compression as an indicator of upright posture. When the rat stands freely in the center of the enclosure,

the body length visually shortens as the nose and tail-base positions converge. In extreme cases, where the rat is fully upright, these two keypoints may fully overlap. To quantify this behavior, a global average of the Euclidean distance between the nose and tail-base was computed across all frames in the video. Then, on a per-frame basis, if the *nose* to *tail-base* distance (i.e., size of the rat) was at most a threshold t% of the global average, the frame was classified as `Standing`. A lower value of the threshold t means a smaller rat size and therefore a more upright position. A higher value of the threshold is more permissive in terms of standing, i.e. allows less strict upright positions to be labeled as `Standing`.

For the classification of `Object Interaction`, only the nose position was required. Prior to analysis, the user needed to annotate each object of interest within the experimental arena using bounding boxes (or circles). Given that objects in our dataset do not change positions within a given video, this was done manually by extracting the first frame of each video and using the LabelMe annotation tool (Russell et al., 2008) to draw bounding boxes/circles around each object. Once both the nose positions and object boundaries were established, object interaction was determined by evaluating whether the (x, y) coordinates of the nose fall within the rectangular boundary of any annotated object. If the nose position lies within a defined object region, that frame is labeled as `Object Interaction`, and the specific object the rat is interacting with is also identified based on the bounding box coordinates of the object.

The classification of `Freezing` and `Resting` behaviors was based on duration-dependent immobility rules applied to tracked keypoints across a sequence of frames. For a frame sequence to be labeled as either behavior, the following criteria had to be sustained for a minimum of one second, equivalent to 60 consecutive frames. `Resting` was defined as the absence of detectable movement in the *tail-base* over the 60-frame window. Importantly, this rule allows for head or nose motion, such as sniffing, which is what distinguishes resting from freezing. In contrast, `Freezing` was defined more stringently as the absence of movement in both the *nose* and *tail-base* over the same duration, reflecting a more complete state of immobility.

5.4 YOLOv11 models for frame-level behavior classification

We considered two different training approaches for NORT behavior classification: 1) an approach where we trained models to identify specific object interactions (e.g., `Upper` or `Lower` in 2-object videos, or `Object1` to `Object5` in 5-object videos), and 2) an approach where we trained models to identify a generic `Object Interaction` behavior in a combined set of 2-object and 5-object videos, and subsequently used the keypoints and the object bounding boxes with a heuristic-based approach to determine what specific object the rat has interacted with. In both cases, the models were also trained to identify `Standing` and `Other` behaviors, in addition to the specific or generic object interaction behaviors.

One main advantage of the specific-object models is that they require a simpler inference pipeline (without the need for keypoints and postprocessing to find the object's identity). In addition, direct supervision on object identity can potentially yield higher precision. However, one disadvantage of this approach is that it can lead to data fragmentation and class imbalance, reducing effective sample size per class. At the same time, the specific-object approach has poor scalability, as any variations of the arena with the number or arrangement of the objects requires re-labeling and re-training the model, and thus has reduced robustness and generalization to unseen arenas or objects due to overfitting to setup-specific cues (materials, shapes, lighting).

As opposed to that, in the generic-object approach, the model is trained on a pooled, larger, and more diverse dataset, improving representation learning, reducing variance and mitigating the risk of overfitting (Hatamian et al., 2025). The visual characteristics of a rat interacting with an object—such as the proximity of its nose to the object's surface, its posture, and head orientation—are fundamentally similar regardless of whether there are two or five objects in the arena. Thus, this approach is naturally scalable to any number of objects and arrangements as the object identity is derived geometrically (e.g., nearest object to the nose keypoints). In effect, this ensures better transfer to new arenas and

object layouts because the model learns object-agnostic interaction patterns, and also simpler model maintenance as one model serves multiple experimental configurations. However, these advantages come at the cost of a post-hoc mapping from keypoints to object boxes, introducing an additional source of error (i.e., error propagation from pose to final label), and also a slightly more complex inference pipeline and evaluation protocol.

For models trained to identify generic object interactions, we converted the `Object Interaction` predictions into object-specific interactions via a simple pose-geometry heuristic. Similar to what was done in the heuristic approach, we leveraged the YOLOv11 Pose model to obtain the rat's keypoints per frame, specifically, *nose* and *tail-base* coordinates. Subsequently, to find the identity/location of the object that the rat has interacted with, the reference object annotations were used to compute fixed object midpoints. Finally, for each frame, we assigned the interacting object as the object corresponding to the nearest midpoint to the nose. This post-hoc step converts class-agnostic `Object Interaction` predictions into object-specific interactions while preserving the benefits of training a single, generalist model.

We evaluated and compared these two approaches to identify the one that works best for our task, considering that we aimed to minimize the manual annotation effort, and thus have only a limited number of annotated training, validation and test videos, as can be seen from Table 5.1.

All classification models leveraged a pre-trained YOLOv11 backbone, which has been optimized on large-scale image datasets to extract rich, hierarchical features from input frames. This transfer learning approach accelerates convergence and enhances performance on our specialized NORT dataset. The core architecture follows the established three-stage design of a *Backbone* for feature extraction, a *Neck* for multi-scale feature fusion, and a task-specific *Classification Head* for generating frame-level predictions (Figure 2.1). The classification head of the generic-object models has three output neurons corresponding to behavior labels `Object Interaction`, `Standing`, and `Other`. The classification head of the specific-object models for 2-object videos has four output neurons corresponding to the behavior labels `Upper`, `Lower`, `Standing`, and `Other`, while the head for the 5-object videos

has seven output neurons corresponding to `Object1` to `Object5`, `Standing`, and `Other`.

We utilized and compared the full range of YOLOv11 classification model sizes - nano (`n`), small (`s`), medium (`m`), large (`l`), and extra-large (`x`). The selection of hyperparameters was guided by the established default configurations provided within the Ultralytics framework. All models were trained using the AdamW optimizer, known for its effective and stable convergence, with a batch size of 16 for a maximum of 50 epochs. To prevent overfitting we implemented an early stopping mechanism that halts training if validation performance does not improve for 7 consecutive epochs. Training was conducted with an input image size of 224x224 pixels, and mixed-precision training was enabled to optimize computational efficiency.

5.5 Evaluation metrics for NORT behavior classification

We evaluated the NORT behavior classification models using standard classification metrics that quantify both per-class performance and overall robustness across behavioral categories. Specifically, we report precision, recall, and F1-score for each behavior class. Precision measures the proportion of predicted instances of a behavior that are correct, recall measures the proportion of ground-truth instances of a behavior that are correctly identified, and F1-score provides the harmonic mean of precision and recall.

For a given behavior class, precision, recall, and F1-score are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where (TP), (FP), and (FN) denote true positives, false positives, and false negatives, re-

spectively.

Because the behavior classes are imbalanced, we also report macro-averaged precision, recall, and F1-score. Macro-averaging computes each metric independently for each behavior class and then averages across classes, giving equal weight to each class regardless of its frequency. This is important for NORT behavior annotation because some behaviors, such as `Object Interaction` or `Standing`, may occur less frequently than `Other`, but are central to the behavioral analyses.

For the three-class evaluation, all methods were compared using the behavior classes `Standing`, `Object Interaction`, and `Other`. For the object-specific evaluation, predictions were evaluated using the corresponding object-interaction labels: `Upper` and `Lower` for 2-object videos, and `Object 1` through `Object 5` for 5-object videos, together with `Standing` and `Other`. When comparing specific-object models with the generic-object model in the three-class setting, object-specific interaction labels were collapsed into the broader `Object Interaction` category.

In addition to numerical metrics, we use confusion matrices to visualize class-wise prediction patterns and identify systematic confusions among behaviors. These matrices are particularly useful for identifying errors near behavioral transitions, such as transitions between `Object Interaction` and `Other` or between `Standing` and `Other`. Together, the per-class metrics, macro-averaged metrics, and confusion matrices provide a comprehensive evaluation of both behavior-specific discrimination and overall model generalization.

Chapter 6

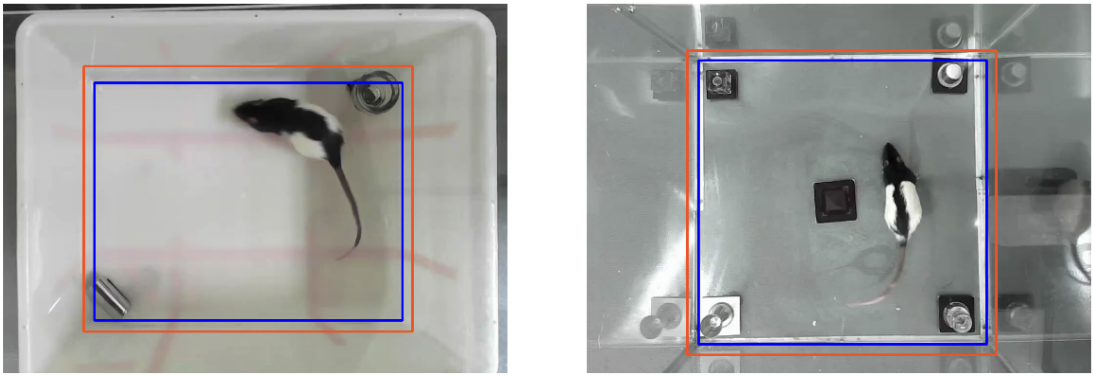
Validation-based optimization for NORT models

To select the final behavior annotation pipeline, we conducted a series of validation-based optimization experiments. These experiments were performed using the validation videos and were used to make three sets of decisions: (1) selecting the hyperparameters for the keypoint-based heuristic approach, (2) selecting the appropriate YOLOv11 classification model size for the combined and separately trained models, and (3) determining whether post-processing improved the frame-level predictions. The final test videos were reserved for the evaluation reported in subsequent sections.

6.1 Hyperparameter optimization for the heuristic-based approach

To identify the optimal configuration for the heuristic-based behavior classification approach, we first conducted hyperparameter fine-tuning experiments using the validation videos. Because the heuristic rules are applied sequentially, we optimized the parameters for `Standing` first and then optimized the parameters for `Object Interaction`.

For `Standing`, we evaluated combinations of arena bounding-box size, specifically tight



(a) *B06_G3:* 2-Object Video

(b) *B06_G3:* 5-Object Video

Figure 6.1: *Sample arena bounding boxes from a 2-object video and from a 5-object video, respectively. Two types of bounding boxes are shown for each type of video: tight bounding boxes (shown in blue) and loose bounding boxes (shown in orange).*

versus loose bounding boxes, as shown in Figure 6.1, and a keypoint-derived *rat body compression* threshold of 70%, 75%, or 80%. The compression threshold was used as a proxy for upright posture, with lower values corresponding to a more compressed body shape and therefore a stronger standing cue. Table 6.1 summarizes precision, recall, and F1 performance across the three validation videos and their averages. The results show a clear precision–recall trade-off: tighter arena bounding boxes generally increased recall but reduced precision, whereas looser arena bounding boxes improved precision at the cost of some missed detections. Similarly, increasing the compression threshold made the detector more permissive, increasing recall while slightly reducing precision. Among all tested configurations, the loose bounding box with a 75% compression threshold achieved the best balance between precision (0.809) and recall (0.817), resulting in the highest average F1 score (0.811). This configuration was therefore selected for the `Standing` heuristic in subsequent analyses.

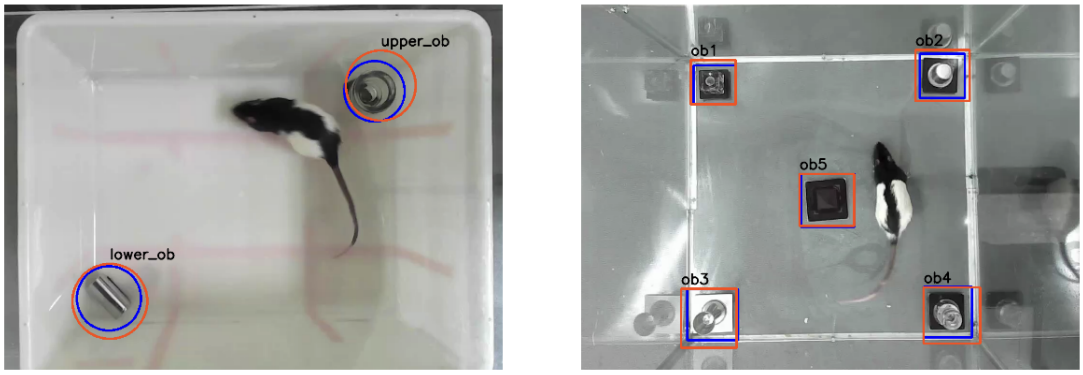
For `Object Interaction`, we evaluated two spatial rules based on object bounding regions, as shown in Figure 6.2. The first rule used a tight object bounding box and classified a frame as an interaction when the rat's nose was within 1 cm of the tight object boundary. The second rule used a loose object bounding box and classified a frame as an interaction when the rat's nose was inside the loose bounding region.

Table 6.2 summarizes validation performance for these two rules in terms of precision,

Video (#obj)	Arena Box/Compression	Standing		
		Pr	Re	F1
B8.13_G3 (2)	Tight bounding box, 70%	0.535	0.881	0.666
	Tight bounding box, 75%	0.523	0.918	0.667
	Tight bounding box, 80%	0.518	**0.951**	0.670
	Loose bounding box, 70%	**0.764**	0.796	0.780
	Loose bounding box, 75%	0.726	0.838	0.778
	Loose bounding box, 80%	0.713	0.891	**0.792**
B8.02_G2 (2)	Tight bounding box, 70%	0.677	0.854	0.755
	Tight bounding box, 75%	0.670	0.884	0.763
	Tight bounding box, 80%	0.652	**0.915**	0.762
	Loose bounding box, 70%	**0.856**	0.787	0.820
	Loose bounding box, 75%	0.837	0.833	**0.835**
	Loose bounding box, 80%	0.800	0.866	0.832
B6.04_B3 (5)	Tight bounding box, 70%	0.422	0.817	0.557
	Tight bounding box, 75%	0.429	0.872	0.575
	Tight bounding box, 80%	0.407	**0.889**	0.558
	Loose bounding box, 70%	**0.907**	0.726	0.806
	Loose bounding box, 75%	0.865	0.781	**0.821**
	Loose bounding box, 80%	0.758	0.803	0.780
Avg (3 videos)	Tight bounding box, 70%	0.545	0.851	0.659
	Tight bounding box, 75%	0.541	0.891	0.668
	Tight bounding box, 80%	0.526	**0.918**	0.663
	Loose bounding box, 70%	**0.842**	0.770	0.802
	Loose bounding box, 75%	0.809	0.817	**0.811**
	Loose bounding box, 80%	0.757	0.853	0.801

Table 6.1: *Validation performance of the heuristic approach for the* `Standing` *behavior. For each group, including the average group, the best performance for each metric is highlighted in* **bold** *font. It can be seen that precision is significantly better for the loose arena bounding box, as compared to the tighter arena, although the recall decreases when using the loose bounding box. On the other hand, precision decreases when compression threshold increases (more permissive, less upright), while the recall increases. Overall, the best F1 score is obtained with a loose bounding box and a compression threshold of 75% (highlighted in blue). This configuration will be used in the heuristic-based approach.*

recall, and F1 score. Across the validation videos, the loose-box rule produced stronger overall performance. Averaged over the three validation videos, the loose-box rule achieved a higher macro F1 score than the tight-box rule (0.846 versus 0.798). In particular, recall for `Object Interaction` increased from 0.730 with the tight-box rule to 0.883 with the loose-



(a) *B06_G3:* 2-Object Video

(b) *B06_G3:* 5-Object Video

Figure 6.2: *Sample object bounding circles/boxes from a 2-object video and from a 5-object video, respectively. Two types of bounding circles/boxes are shown for each type of video: tight bounding circles/boxes (shown in color) and loose bounding circles/boxes (shown in orange).*

box rule, while precision decreased only slightly from 0.897 to 0.869. This improvement suggests that allowing greater spatial flexibility around object boundaries better captures natural exploratory behavior and reduces missed interactions caused by variation in posture, object approach angle, and keypoint localization noise. The loose-box rule was therefore selected for the `Object Interaction` heuristic.

Video (#obj)	Obj. Int. Box/Rule	Standing			Object Interaction			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B8.13_G3 (2)	Tight BB, within 1cm from BB	0.626	**0.884**	0.733	**0.881**	0.680	0.768	**0.921**	0.856	0.887	0.809	0.807	0.796
	Loose BB, nose within BB	**0.726**	0.838	**0.778**	0.864	**0.823**	**0.843**	0.915	**0.880**	**0.897**	**0.835**	**0.847**	**0.839**
B8.02_G2 (2)	Tight, within 1cm from BB	0.773	**0.869**	0.818	**0.925**	0.771	0.841	**0.792**	0.763	0.778	0.830	0.801	0.812
	Loose BB, nose within BB	**0.837**	0.833	**0.835**	0.888	**0.895**	**0.891**	0.787	**0.788**	**0.787**	**0.837**	**0.839**	**0.838**
B6.04_B3 (5)	Tight, within 1cm from BB	0.653	**0.849**	0.738	**0.884**	0.739	0.805	0.787	**0.850**	0.817	0.775	0.813	0.787
	Loose BB, nose within BB	**0.865**	0.781	**0.821**	0.854	**0.930**	**0.890**	**0.904**	0.844	**0.873**	**0.874**	**0.852**	**0.861**
Avg (3 videos)	Tight BB, within 1cm from BB	0.684	**0.867**	0.763	**0.897**	0.730	0.805	0.833	0.823	0.827	0.805	0.807	0.798
	Loose BB, nose within BB	**0.809**	0.817	**0.811**	0.869	**0.883**	**0.875**	**0.869**	**0.837**	**0.852**	**0.849**	**0.846**	**0.846**

Table 6.2: *Validation performance of the `Object Interaction` heuristic. Two rules are compared on each model. Averages over the three validation videos are also shown for each rule. For each video group and also for their average group, best performance for each rule in that group is highlighted in* **bold** *font. The best overall performance is obtained with the rule that uses loose bounding boxes and checks if the nose in within the box (highlighted in blue font).*

The final heuristic-based configuration therefore used a loose arena bounding box with a 75% body-compression threshold for `Standing`, followed by the loose object-bounding-region rule for `Object Interaction`. This configuration was used in the heuristic-only evaluation

and also served as the spatial reasoning component for assigning specific object identities to generic object-interaction predictions produced by the combined YOLOv11 model.

6.2 Hyperparameter optimization for YOLOv11 classification models

We next conducted hyperparameter fine-tuning experiments to identify the optimal YOLOv11 classification model configurations. This analysis focused on the effect of model size on frame-level behavior classification performance. We evaluated the full range of YOLOv11 classification model sizes: nano, small, medium, large, and extra-large. For each model size, we trained two types of models: generic object-interaction models trained on the combined set of 2-object and 5-object videos, and specific object-interaction models trained separately on the 2-object and 5-object videos.

All models were evaluated on the validation videos in terms of their ability to predict `Standing`, `Object Interaction`, and `Other`. For the specific object-interaction models, the output labels included object-specific interaction classes. However, for direct comparison with the generic object-interaction models, these object-specific labels were collapsed into the broader `Object Interaction` category. This allowed us to compare model sizes and training strategies using a shared three-class evaluation.

Tables 6.3 and 6.4 summarize the validation results for models trained on the combined 2-object/5-object dataset and on the separate 2-object and 5-object datasets, respectively. The results show that different model sizes performed best for different videos or metrics. However, when trained on the combined dataset, YOLOv11-`large` achieved the best overall macro-averaged performance and was therefore selected as the generic object-interaction model. When trained separately on the 2-object and 5-object datasets, YOLOv11-`medium` achieved the best overall performance and was therefore selected for the specific object-interaction models.

The comparison also suggests that the generic object-interaction models trained on the

Video (#obj)	Model	Standing			Object Interaction			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B8.13_G3 (2)	nano	**0.960**	0.944	0.952	0.910	0.915	0.912	0.957	0.961	0.959	0.942	0.940	0.941
	small	0.943	**0.977**	**0.960**	0.901	0.933	0.916	**0.972**	0.950	0.961	0.938	**0.953**	0.946
	medium	0.951	0.969	**0.960**	0.930	0.916	0.923	0.965	0.964	**0.964**	0.949	0.949	**0.949**
	large	0.901	0.958	0.929	0.907	**0.951**	0.928	0.971	0.936	0.953	0.926	0.948	0.937
	xlarge	0.943	0.931	0.937	**0.952**	0.938	**0.945**	0.959	**0.967**	0.963	**0.951**	0.945	0.948
B8.02_G2 (2)	nano	0.957	0.915	0.935	0.877	**0.968**	0.920	0.891	0.875	0.883	0.908	0.919	0.913
	small	0.935	0.950	0.942	0.912	0.959	0.935	0.912	0.862	0.887	0.920	0.924	0.921
	medium	0.914	0.954	0.934	0.904	0.945	0.924	0.918	0.840	0.877	0.912	0.913	0.912
	large	0.927	**0.964**	0.945	**0.927**	0.957	0.942	**0.934**	0.867	0.899	0.929	0.929	0.929
	xlarge	**0.967**	0.937	**0.952**	0.920	0.966	**0.943**	0.915	**0.918**	**0.917**	**0.934**	**0.940**	**0.937**
B6.04_B3 (5)	nano	0.735	**0.878**	0.800	0.913	0.762	0.831	0.776	0.868	0.820	0.808	0.836	0.817
	small	0.853	0.840	0.846	0.892	0.847	0.869	0.812	0.862	0.836	0.852	0.850	0.850
	medium	0.849	0.827	0.838	**0.933**	0.774	0.846	0.756	**0.908**	0.825	0.846	0.837	0.836
	large	0.827	0.861	0.843	0.917	**0.868**	**0.892**	**0.845**	0.884	**0.864**	**0.863**	**0.871**	**0.866**
	xlarge	**0.882**	0.820	**0.850**	0.918	0.808	0.859	0.773	0.897	0.830	0.858	0.842	0.847
Avg (3 videos)	nano	0.884	0.912	0.896	0.900	0.882	0.888	0.875	0.901	0.887	0.886	0.898	0.890
	small	0.910	0.922	**0.916**	0.902	0.913	0.907	0.899	0.891	0.895	0.903	0.909	0.906
	medium	0.905	0.917	0.911	0.922	0.878	0.898	0.880	0.904	0.889	0.902	0.900	0.899
	large	0.885	**0.928**	0.906	0.917	**0.925**	**0.921**	**0.917**	0.896	**0.905**	0.906	**0.916**	**0.911**
	xlarge	**0.931**	0.896	0.913	**0.930**	0.904	0.916	0.882	**0.927**	0.903	**0.914**	0.909	**0.911**

Table 6.3: *Validation performance of YOLOv11 classification models trained on the combined set of 2-object and 5-object videos. Various model sizes (including nano, small, medium, large and xlarge) are evaluated on the validation set (which includes two 2-object videos and one 5-object video). Performance metrics are shown for each behavior separately and also macro-averaged over all behaviors. For each video group and also for their average group, best performance for each metric in that group is highlighted in* **bold** *font. While no model has best performance for all validation videos and metrics, the large version of YOLOv11 has overall best performance (highlighted in blue color), when trained on the combined set of 2-object and 5-object videos.*

combined dataset performed better overall than the specific object-interaction models trained separately. This result is consistent with the fact that the combined dataset provides more training examples and allows the model to learn object-interaction patterns across different arena configurations. Based on these validation results, we use YOLOv11-large as the combined generic-object model and YOLOv11-`medium` as the separately trained specific-object model in the remaining experiments.

6.3 Post-processing optimization

After selecting the YOLOv11 model configurations, we evaluated whether post-processing could further improve frame-level prediction performance. Specifically, we investigated a sequential two-step post-processing procedure consisting of *Interaction Relabeling* (IR) fol-

Video (#obj)	Model	Standing			Object Interaction			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B8.13_G3 (2)	nano	0.831	0.771	0.800	**0.895**	0.519	0.657	0.810	0.929	0.865	0.845	0.740	0.774
	small	0.846	0.856	0.851	0.898	0.450	0.599	0.818	0.934	0.872	0.854	0.747	0.774
	medium	**0.943**	0.861	**0.900**	0.861	0.722	0.785	0.883	**0.951**	**0.916**	**0.896**	0.845	**0.867**
	large	0.860	**0.908**	0.883	0.840	0.764	0.800	0.902	0.909	0.905	0.867	**0.860**	0.863
	xlarge	0.828	0.798	0.812	0.756	**0.882**	**0.814**	**0.911**	0.877	0.894	0.832	0.852	0.840
B8.02_G2 (2)	nano	**0.982**	0.757	0.855	0.866	0.705	0.777	0.657	**0.921**	0.767	0.835	0.794	0.800
	small	0.905	0.917	0.911	**0.950**	0.787	0.861	0.796	0.877	**0.834**	**0.884**	0.860	0.869
	medium	0.905	**0.937**	**0.921**	0.828	0.947	**0.884**	**0.902**	0.775	**0.834**	0.878	**0.886**	**0.879**
	large	0.952	0.881	0.915	0.757	0.927	0.833	0.823	0.771	0.796	0.844	0.859	0.848
	xlarge	0.933	0.866	0.898	0.684	**0.989**	0.809	0.832	0.651	0.730	0.816	0.835	0.812
B6.04_B3 (5)	nano	0.639	0.887	0.743	0.853	0.697	0.767	0.740	0.792	0.765	0.744	0.792	0.758
	small	0.816	0.837	0.826	0.888	0.684	0.773	0.687	0.860	0.764	0.797	0.794	0.787
	medium	0.720	**0.898**	0.799	0.837	**0.838**	**0.838**	**0.836**	0.762	0.798	0.798	**0.833**	0.811
	large	0.797	0.828	0.812	**0.899**	0.768	0.828	0.752	**0.866**	0.805	0.816	0.821	0.815
	xlarge	**0.819**	0.852	**0.835**	0.883	0.783	0.830	0.769	0.857	**0.811**	**0.823**	0.831	**0.825**
Avg (3 videos)	nano	0.817	0.805	0.799	0.871	0.640	0.734	0.736	0.881	0.799	0.808	0.775	0.777
	small	0.856	0.870	0.863	**0.912**	0.640	0.744	0.767	**0.890**	0.823	0.845	0.800	0.810
	medium	0.856	**0.899**	**0.873**	0.842	0.836	**0.836**	**0.874**	0.829	**0.849**	**0.857**	**0.855**	**0.852**
	large	**0.870**	0.872	0.870	0.832	0.820	0.820	0.826	0.849	0.835	0.842	0.847	0.842
	xlarge	0.860	0.839	0.848	0.774	**0.885**	0.818	0.837	0.795	0.812	0.824	0.839	0.826

Table 6.4: *Validation performance of YOLOv11 classification models trained separately on 2-object and 5-object videos, respectively. Various model sizes (including nano, small, medium, large and xlarge) are evaluated on the validation set (which includes two 2-object videos and one 5-object video). Performance metrics are shown for each behavior separately and also macro-averaged over all behaviors. For each video group and also for their average group, best performance for each metric in that group is highlighted in* **bold** *font. While no model has best performance for all validation videos and metrics, the medium version of YOLOv11 has overall best performance (highlighted in blue color), when trained separately on the 2-object videos and 5-object videos.*

lowed by *Temporal Smoothing* (TS).

Interaction Relabeling was motivated by visual inspection of the model predictions, which suggested that some frames corresponding to `Object Interaction` were incorrectly labeled as `Other`. To address this issue, we revisited frames predicted as `Other` and applied the selected `Object Interaction` heuristic. If the rat's nose was inside the loose object bounding region, the frame was relabeled as `Object Interaction`.

Temporal Smoothing was used to reduce isolated frame-level label fluctuations. We applied a majority-vote filter over a temporal sliding window of size (2k+1) centered at each frame, excluding the current frame. If any label appeared more than (k+1) times within the window and the current label differed from the dominant label, the current frame label was replaced with the dominant label. We used (k=15) for both 2-object and 5-object videos.

Tables 6.5 and 6.6 compare raw model predictions with the sequential IR and IR+TS

post-processing results for the combined generic object-interaction model and the separately trained specific object-interaction models, respectively. For the combined YOLOv11-`large` model, post-processing did not improve overall performance; the raw model achieved the best average macro F1 score. Therefore, the final test results for the combined YOLOv11-`large` model are reported without post-processing. In contrast, for the separately trained YOLOv11-`medium` models, the IR+TS post-processing procedure produced a small improvement in average macro F1 score, from 0.851 for the raw model to 0.856 after IR+TS. Therefore, the final test results for the separately trained YOLOv11-`medium` models are reported using IR+TS post-processed predictions.

Video (#obj)	Post-Proc	Standing			Object Interaction			Other			Macro_avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B8.13_G3 (2)	-	0.901	0.959	0.929	**0.908**	0.951	**0.929**	0.971	**0.937**	**0.954**	**0.927**	**0.949**	**0.937**
	IR	0.775	0.962	0.859	0.876	0.984	0.927	0.982	0.867	0.921	0.878	0.938	0.902
	IR+TS	0.776	0.966	0.861	0.872	0.988	0.926	0.985	0.865	0.921	0.877	0.940	0.903
B8.02_G2 (2)	-	**0.927**	0.965	0.945	**0.927**	0.957	**0.942**	0.934	**0.868**	**0.900**	**0.929**	**0.930**	**0.929**
	IR	0.846	0.975	0.906	0.902	0.974	0.937	0.954	0.727	0.825	0.901	0.892	0.889
	IR+TS	0.845	**0.979**	0.907	0.908	0.978	0.941	0.961	0.727	0.828	0.904	0.895	0.892
B6.04_B3 (5)	-	0.828	0.861	0.844	**0.918**	0.868	0.893	0.845	**0.885**	0.865	0.864	0.871	0.867
	IR	0.795	0.885	0.838	0.862	0.966	0.911	0.969	0.797	0.874	0.875	0.883	0.874
	IR+TS	0.801	**0.891**	0.843	0.865	**0.970**	**0.915**	**0.973**	0.800	**0.878**	**0.880**	**0.887**	**0.879**
Avg (3 videos)	-	**0.885**	0.928	**0.906**	**0.918**	0.925	0.921	0.917	**0.897**	**0.906**	**0.907**	**0.917**	**0.911**
	IR	0.805	0.941	0.868	0.880	0.975	0.925	0.968	0.797	0.873	0.885	0.904	0.888
	IR+TS	0.807	**0.945**	0.870	0.882	**0.979**	**0.927**	**0.973**	0.797	0.876	0.887	0.907	0.891

Table 6.5: *Post-processing validation performance of the YOLOv11-*`large` *classification model trained on the combined set of 2-object and 5-object videos. The post-processing consists of two steps applied sequentially. In the first step (IR), the* ***`Other`*** *behavior is relabeled as* ***`Object Interaction`****, if an* ***`Object Interaction`*** *rule applies (i.e., loose box and nose within box). Subsequently, in the second step (TS), a sliding window is used to smoothen the labels within a window based on the majority label in that window. The results of the sequential post-processing steps are compared with those of the raw model without any post-processing (-). Performance metrics are shown for each behavior separately and also macro-averaged over all behaviors. For each video group and also for the average group, best performance for each metric in that group is highlighted in* **bold** *font. The average comparison shows that the two-step IR+TS post-processing slightly helps in some cases, but the raw model has the best overall performance (highlighted in blue color).*

Video (#obj)	Post-Proc	Standing			Object Interaction			Other			Macro_avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B8.13_G3 (2)	-	**0.944**	0.862	**0.901**	**0.862**	0.722	0.786	0.884	**0.952**	**0.916**	**0.896**	0.845	0.868
	IR	0.754	0.931	0.834	0.834	0.926	0.877	**0.958**	0.851	0.901	0.849	0.903	**0.871**
	IR+TS	0.753	**0.932**	0.833	0.834	**0.931**	**0.880**	**0.958**	0.849	0.901	0.849	**0.904**	**0.871**
B8.02_G2 (2)	-	**0.905**	0.938	**0.921**	**0.829**	0.948	0.884	0.903	**0.775**	**0.834**	**0.879**	**0.887**	**0.880**
	IR	0.819	0.965	0.886	0.821	0.977	**0.892**	0.968	0.633	0.765	0.869	0.859	0.848
	IR+TS	0.821	**0.972**	0.890	0.818	**0.979**	0.891	**0.978**	0.629	0.765	0.872	0.860	0.849
B6.04_B3 (5)	-	0.720	0.898	0.799	**0.838**	0.839	0.838	0.837	**0.763**	0.798	0.798	0.833	0.812
	IR	0.709	0.936	0.807	0.826	0.928	0.874	0.960	0.718	0.821	0.832	0.861	0.834
	IR+TS	**0.740**	**0.944**	**0.830**	0.834	**0.948**	**0.887**	**0.966**	0.722	**0.826**	**0.847**	**0.871**	**0.848**
Avg (3 videos)	-	**0.856**	0.899	**0.874**	**0.843**	0.836	0.836	0.875	**0.830**	**0.849**	**0.858**	0.855	0.851
	IR	0.761	0.944	0.842	0.827	0.944	0.881	0.962	0.734	0.829	0.850	0.874	0.851
	IR+TS	0.771	**0.950**	0.851	0.829	**0.953**	**0.886**	**0.967**	0.733	0.831	0.856	**0.878**	**0.856**

Table 6.6: *Post-processing validation performance of the YOLOv11-*`medium` *classification models trained independently on the 2-object and 5-object videos, respectively. The post-processing consists of two steps applied sequentially. In the first step (IR), the* `Other` *behavior is relabeled as* `Object Interaction`*, if an* `Object Interaction` *rule applies (i.e., loose box and nose within box). Subsequently, in the second step (TS), a sliding window is used to smoothen the labels within a window based on the majority label in that window. The results of the sequential post-processing steps are compared with those of the raw model without any post-processing (-). Performance metrics are shown for each behavior separately and also macro-averaged over all behaviors. For each video group and also for the average group, best performance for each metric in that group is highlighted in* **bold** *font. The average comparison shows that the two-step IR+TS post-processing has slightly better overall performance (highlighted in blue color) as compared to the other two variants, improving recall at the expense of precision for the* `Standing` *and* `Object Interaction` *behaviors.*

6.4 Human agreement comparison

To contextualize model performance relative to human annotation variability, we compared the selected combined YOLOv11-`large` model with student annotations and domain-expert annotations on representative validation videos. In this comparison, the domain-expert annotations were treated as the reference labels, and both the student annotations and model predictions were evaluated against them. This analysis provides an estimate of how close the model performance is to human-level annotation under the same frame-level behavior definitions.

Table 6.7 compares student and model performance on the 2-object validation video B8.13_G3. Overall, the model achieved performance comparable to the student annotations. The model had slightly higher macro-averaged F1 than the student annotations, with particularly strong performance on the object-interaction classes and `Other`. The Cohen's kappa

coefficient between the student and domain-expert annotations was 0.872, indicating strong agreement between human annotators.

Behavior	Student Performance			Model Performance		
	Pr	Re	F1	Pr	Re	F1
Standing	0.912	0.957	0.934	0.901	0.959	0.929
Lower	0.771	0.954	0.853	0.902	0.944	0.923
Upper	0.914	0.919	0.910	0.917	0.964	0.940
Other	0.969	0.910	0.939	0.971	0.937	0.954
Macro-avg	0.892	0.935	0.909	0.923	0.951	0.936

Table 6.7: *Human Agreement Results versus Combined YOLOv11-*`large` *Model Results on 2-object validation video B8.13_G3. The metrics are calculated by comparing student/model annotations with the domain expert annotations used as ground truth in this work. Overall, the metrics of the student and the model are very similar. The model has better F1 and precision for* ***`Lower Object`*** *but slightly lower recall, comparable performance on* ***`Standing`*** *and slightly better performance on* ***`Upper Object`****. The Cohen Kappa coefficient for the agreement between the student and the domain expert annotations is* 0.872*.*

Table 6.8 shows the corresponding comparison for the 5-object validation video B6.04_B3. In this setting, the model again achieved performance comparable to student-level annotation, although the relative strengths differed across behaviors. The model performed well for `Standing` and several object-interaction classes, while the student annotations achieved higher recall for some object-specific interactions. The Cohen's kappa coefficient between the student and domain-expert annotations was 0.863, again indicating strong human-human agreement.

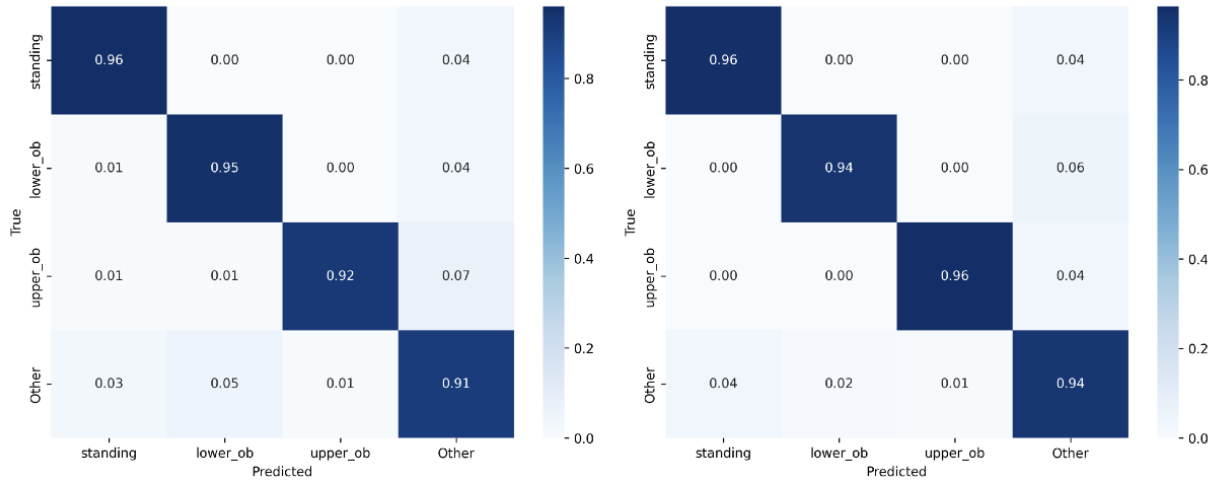
Figure 6.3 shows normalized confusion matrices comparing student and model annotations against the domain-expert reference labels for the same validation videos. These matrices indicate that both the student and the model make most errors near behavioral transitions, such as when the rat changes from `Object Interaction` to `Other`, from `Standing` to `Other`, or from `Standing` to `Object Interaction`. These transition frames are inherently difficult because the behavioral state may change gradually rather than instantaneously, and even human annotators may disagree about the exact frame at which one behavior ends and another begins.

Taken together, these results suggest that the selected combined YOLOv11-`large` model

Behavior	Student Performance			Model Performance		
	Pr	Re	F1	Pr	Re	F1
Standing	0.873	0.782	0.825	0.828	0.861	0.844
Object 1	0.843	0.991	0.911	0.898	0.736	0.809
Object 2	0.853	0.983	0.913	0.897	0.903	0.900
Object 3	0.915	0.989	0.950	0.923	0.963	0.942
Object 4	0.885	0.990	0.935	0.956	0.904	0.930
Object 5	0.861	0.961	0.908	0.896	0.845	0.870
Other	0.925	0.825	0.872	0.845	0.885	0.865
Macro-avg	0.879	0.932	0.902	0.892	0.871	0.880

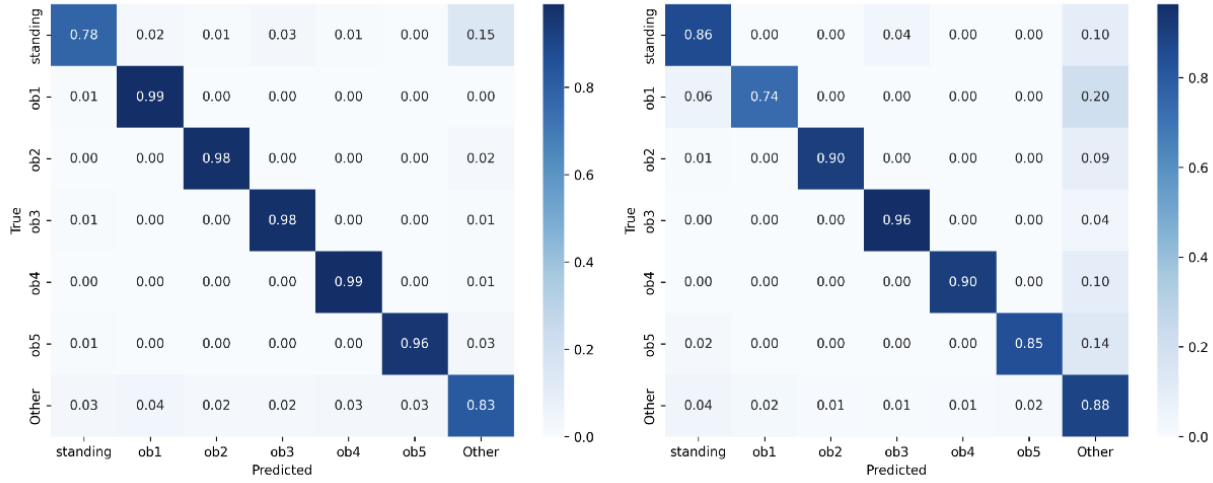
Table 6.8: *Human agreement results versus combined YOLOv11-*`large` *model results on 5-object validation video B6.04_B3. The metrics are calculated by comparing student/model annotations with the domain expert annotations used as ground truth in this work. Overall, the model has better F1 and recall for* `Standing` *but lower precision, and better precision on* `Object Interaction` *but slightly lower F1 and recall. The Cohen Kappa coefficient for the agreement between the student and the domain expert annotations is* 0.863*.*

performs at a level comparable to student human annotation for the NORT frame-level behavior classification task. The remaining errors are concentrated in ambiguous transition frames, which reflects a limitation of frame-level behavioral annotation rather than only a limitation of the model.



(a) *B8.13_G3:* student

(b) *B8.13_G3:* model



(c) *B6.04_B3:* student

(d) *B6.04_B3:* model

Figure 6.3: *Normalized confusion matrices illustrating the classification performance of student/model annotations by comparison with the domain expert annotations. Each row corresponds to one validation video. High diagonal values indicate strong within-class accuracy, while off-diagonal values reveal systematic confusions. For both student/model misclassifications occur near behavioral transitions (e.g.,* `Object Interaction` ↔ `Other`*, Standing* ↔ `Other` *or* `Standing` ↔ `Object Interaction`*), where frame-level ambiguity and annotation inconsistencies are greatest.*

Chapter 7

NORT results

In this chapter, we evaluated the results of the heuristic-based approach by comparison with the YOLOv11 classification models trained from the combined 2-object and 5-object dataset and also individually from the 2-object and 5-object video datasets, respectively. We evaluated all approaches in terms of their ability to predict generic `Object Interaction`, `Standing`, and `Other` behaviors, as well as their ability to accurately predict specific object interactions. Finally, we performed error analysis to understand the limitations of the models.

7.1 Performance results for three-class predictions

Table 7.1 compares three approaches for predicting `Standing`, `Object Interaction`, and `Other` behaviors in NORT videos of rats. The combined model, trained jointly on both 2-object and 5-object datasets, achieved the highest overall performance across nearly all metrics, attaining the best *Macro-avg* precision, recall, and F1-score for both 2-object and 5-object video groups. This result suggests that exposure to a more diverse set of object configurations enabled the model to learn a generalized representation of object-interaction behavior, improving balance between precision and recall across behavioral classes. In contrast, the 2-object and 5-object models, trained specifically on their respective subsets, often achieved slightly higher recall for `Object Interactions` (e.g., 0.988 and 0.955 for 2-object

and 5-object groups, respectively), but at the cost of lower precision and overall F1. The heuristic baseline maintained competitive recall values in some videos but consistently underperformed in F1, reflecting a tendency to over-predict interactions. Overall, the results indicate that the combined model provides the most robust and transferable behavior predictions across varying experimental setups, whereas the specialized models may be advantageous when maximizing recall for interaction detection is a specific experimental priority.

Video (#obj.)	Model	Standing			Object Int.			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B06_G3 (2)	heuristic	0.872	0.838	0.854	0.839	0.649	0.732	0.727	0.850	0.784	0.813	0.779	0.790
	2-object	0.853	**0.979**	0.911	0.904	0.857	0.880	0.892	0.783	0.834	0.883	0.873	0.875
	combined	**0.934**	0.957	**0.946**	**0.937**	**0.958**	**0.947**	**0.931**	**0.896**	**0.913**	**0.934**	**0.937**	**0.935**
F8.16_G2.3 (2)	heuristic	0.793	0.922	0.853	0.682	**0.994**	0.809	0.977	0.873	0.922	0.817	**0.930**	0.861
	2-object	0.714	**0.990**	0.830	0.561	0.864	0.680	0.996	0.804	0.890	0.757	0.886	0.800
	combined	**0.890**	0.970	**0.928**	**0.793**	0.875	**0.832**	**0.983**	**0.942**	**0.962**	**0.889**	0.929	**0.907**
Avg. (2)	heuristic	0.833	0.880	0.854	0.761	0.822	0.771	0.852	0.862	0.853	0.815	0.855	0.826
	2-object	0.784	**0.985**	0.871	0.733	0.861	0.780	0.944	0.794	0.862	0.820	0.880	0.838
	combined	**0.912**	0.964	**0.937**	**0.865**	**0.917**	**0.890**	**0.957**	**0.919**	**0.938**	**0.911**	**0.933**	**0.922**
B10_B2 (5)	heuristic	0.784	0.788	0.786	0.888	**0.944**	**0.915**	0.923	0.888	**0.905**	0.865	0.873	0.869
	5-object	0.738	0.899	0.811	0.855	0.921	0.887	**0.955**	0.836	0.892	0.849	**0.885**	0.863
	combined	**0.914**	**0.818**	**0.864**	**0.933**	0.835	0.881	0.863	**0.948**	0.903	**0.903**	0.867	**0.883**
F1a.09_B3 (5)	heuristic	0.913	0.452	0.604	0.569	0.944	0.710	0.826	0.852	0.839	0.769	0.749	0.718
	5-object	0.886	0.808	0.845	0.592	**0.994**	0.742	**0.945**	0.665	0.781	0.808	0.822	0.789
	combined	**0.985**	**0.849**	**0.912**	**0.781**	0.870	**0.835**	0.861	**0.883**	**0.872**	**0.876**	**0.867**	**0.873**
Avg. (5)	heuristic	0.849	0.620	0.695	0.729	0.944	0.813	0.875	0.870	0.872	0.818	0.811	0.793
	5-object	0.812	**0.854**	0.828	0.724	**0.958**	0.815	**0.950**	0.751	0.837	0.829	0.854	0.826
	combined	**0.950**	0.834	**0.888**	**0.857**	0.853	**0.858**	0.862	**0.916**	**0.916**	**0.890**	**0.868**	**0.887**

Table 7.1: *Performance results on test videos. Three approaches are compared: the heuristic-based approach, a YOLOv11-*`large` *model trained on the combined 2-object and 5-object dataset, and YOLOv11-*`medium-IR+TS` *models trained independently on the 2-object and 5-object videos, respectively. Models are evaluated in terms of their ability to predict* ***`Standing`****, generic* ***`Object Interactions`*** *and* ***`Other`*** *behaviors. For each video group and also for the 2-object and 5-object average groups, best performance for each metric in that group is highlighted in* **bold** *font. For both groups, the best results overall, highlighted in blue are obtained with the combined model.*

7.2 Performance results for specific object interactions

Table 7.2 compares the heuristic approach, the YOLOv11 model trained only on 2-object videos, and the YOLOv11-`large` model trained on the combined 2-object and 5-object dataset in terms of specific object interactions in 2-object test videos, i.e., `Upper Object` and `Lower Object`. Averaged over the two 2-object videos, the combined model yielded

the strongest F1 for both `Lower Object` (0.885) and `Upper Object` (0.888), with balanced precision/recall values (Pr≈0.86-0.87, Re≈0.90-0.92). The 2-object-only model attained good recall (e.g., `Upper Object` Re=0.892) but substantially lower precision (e.g., `Lower Object` Pr=0.684). The heuristic was competitive on recall (especially for `Upper Object` in *F8.16_G2.3*, Re=1.00) but typically traded off precision, leading to lower average F1 as compared to both the combined model and the 2-object model.

Video (#obj.)	Model	Lower Object			Upper Object		
		Pr	Re	F1	Pr	Re	F1
B06_G3 (2)	heuristic	0.770	0.903	0.831	0.899	0.537	0.673
	2-object	0.814	**0.999**	0.897	0.963	0.794	0.871
	combined	**0.927**	0.944	**0.935**	**0.942**	**0.964**	**0.946**
F8.16_G2.3 (2)	heuristic	0.716	**0.989**	0.831	0.644	**1.000**	0.784
	2-object	0.553	0.764	0.642	0.569	0.990	0.723
	combined	**0.803**	0.868	**0.834**	**0.781**	0.884	**0.829**
Average (2)	heuristic	0.743	**0.946**	0.831	0.772	0.769	0.729
	2-object	0.684	0.882	0.770	0.766	0.892	0.797
	combined	**0.865**	0.906	**0.885**	**0.862**	**0.924**	**0.888**

Table 7.2: *Performance results for specific object interactions in the 2-object test videos. Three approaches are compared: the heuristic-based approach, a YOLOv11-*`large` *model trained on the combined 2-object and 5-object dataset, and a model trained independently on the 2-object videos. Models are evaluated in terms of their ability to predict specific object interactions, i.e.,* ***`Lower Object`*** *and* ***`Upper Object`*** *behaviors. For each video group and for the 2-object average group, best performance for each metric is highlighted in* **bold***; ties are bolded in all tied entries.*

Table 7.3 shows performance for specific objects in the 5-object test videos. On average across the two 5-object videos, the combined model again provided the most consistent F1 - better than both the heuristic and the 5-object-only model for `Objects 2-4` (e.g., Obj3 F1=0.919, Obj4 F1=0.903) and comparable with the results of the 5-object-only model on `Object 1` (with F1=0.765 for the 5-object-only model and F1=0.760 for the combined model). Notably, the heuristic was strongest on `Object 5` (Avg F1=0.952 vs. 0.923 for the 5-object model and 0.879 for the combined model), and also better for several objects in *B10_B2* (e.g., `Objects 2, 4 and 5`). These exceptions are consistent with scene-specific geometry: when the rat sits directly on or partially occludes an object, the rule-based can be more robust than appearance cues, whereas the YOLOv11 models occasionally revert to

`Other`. The 5-object-only model showed a similar pattern to the 2-object-only model: recall remained high (often > 0.95) but precision lagged, producing lower F1.

Video (#obj.)	Model	Object 1			Object 2			Object 3			Object 4			Object 5		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
B10_B2 (5)	heuristic	**0.899**	0.779	0.835	0.913	**0.905**	**0.909**	0.832	**0.998**	0.907	0.864	0.953	**0.906**	0.951	**0.967**	**0.959**
	5-object	0.770	**0.989**	0.866	0.854	0.831	0.843	0.817	0.980	0.891	0.817	**0.982**	0.892	**0.983**	0.870	0.923
	combined	0.876	0.872	**0.874**	**0.927**	0.770	0.841	**0.930**	0.969	**0.949**	**0.923**	0.877	0.899	0.978	0.708	0.821
F1a.09_B3 (5)	heuristic	0.421	0.860	0.565	0.399	0.914	0.555	0.693	0.991	0.816	0.642	0.959	0.769	0.900	0.991	**0.944**
	5-object	0.496	**0.999**	0.663	0.414	**0.995**	0.585	0.684	**0.997**	0.812	0.706	**0.979**	0.820	0.856	**0.999**	0.922
	combined	**0.515**	0.867	**0.646**	**0.815**	0.853	**0.834**	**0.832**	0.955	**0.889**	**0.890**	0.922	**0.906**	**0.983**	0.894	0.937
Average (5)	heuristic	0.660	0.820	0.700	0.656	0.910	0.732	0.763	**0.995**	0.862	0.753	0.956	0.838	0.926	**0.979**	**0.952**
	5-object	0.633	**0.994**	**0.765**	0.634	**0.913**	0.714	0.751	0.989	0.852	0.762	**0.981**	0.856	0.920	0.935	0.923
	combined	**0.696**	0.870	0.760	**0.871**	0.812	**0.838**	**0.881**	0.962	**0.919**	**0.907**	0.900	**0.903**	**0.981**	0.801	0.879

Table 7.3: *Performance results for specific object interactions in the 5-object test videos. Three approaches are compared: the heuristic-based approach, a YOLOv11-*`large` *model trained on the combined 2-object and 5-object dataset, and a YOLOv11-*`medium-IR+TS` *model trained on the 5-object videos. Best (maximum) values for each metric within a video group (and within the Average group) are* **bolded**.

Across both settings, the combined YOLOv11 model offered the best overall trade-off between precision and recall and the most stable F1 across objects and videos, while single-subset (2-object and 5-object videos, respectively) models biased toward recall (over-prediction near interaction onsets/offsets). The heuristic excels in specific layouts (notably `Object 5`), where a fixed geometric tolerance may trump appearance features (e.g., cases where the object is occluded by the rat).

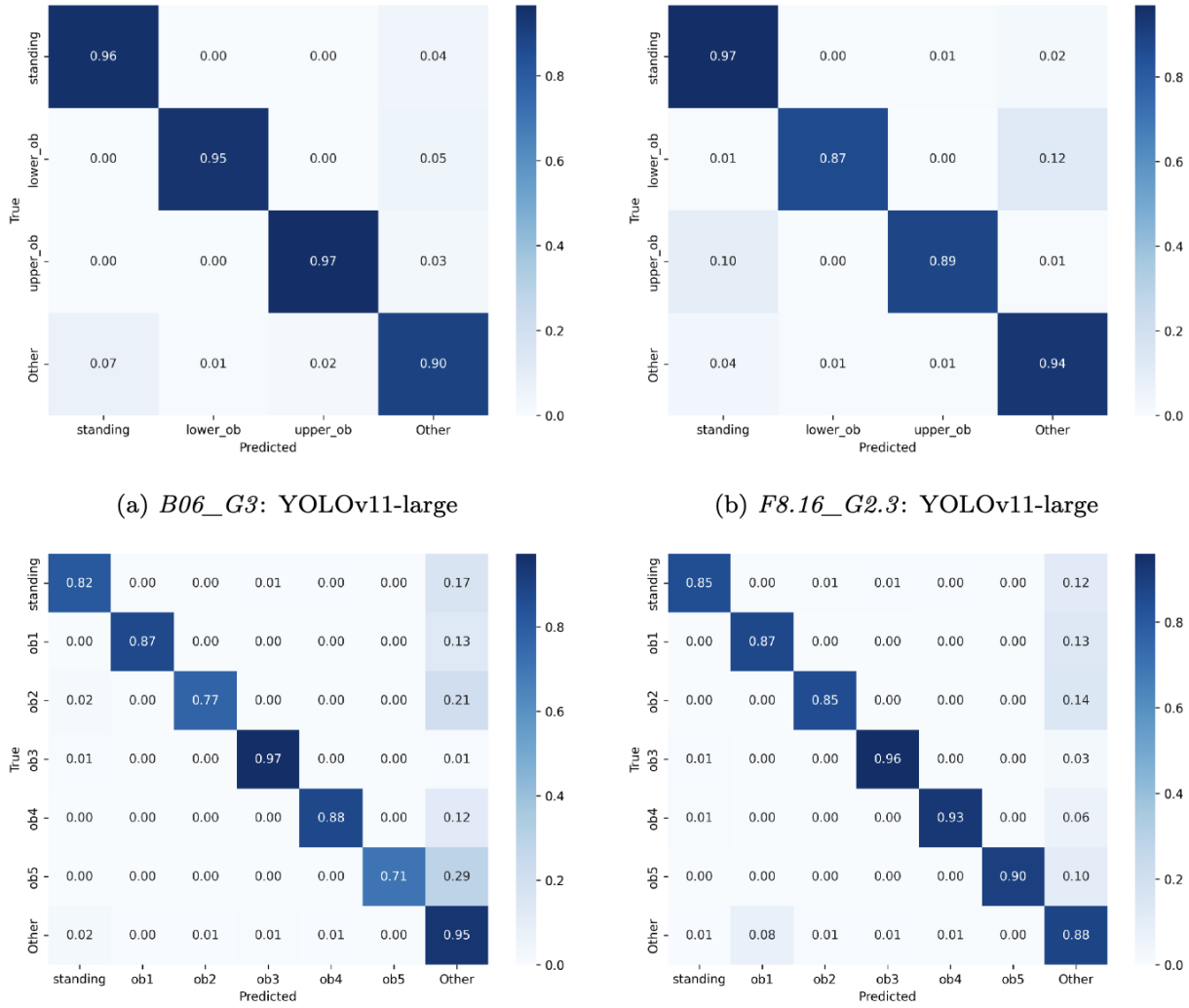
7.3 Error analysis for behavior classification

To gain deeper insight into YOLOv11-`large` model performance and error patterns, we analyzed normalized confusion matrices for each test video, shown in Figure 7.1 (similar confusion matrices for the single-subset YOLOv11-`medium-post` models and for the heuristic approach are shown in Figures 7.2 and 7.3, respectively). The matrices reveal strong diagonal dominance across all behaviors, confirming that the YOLOv11-`large` model accurately capture the main behavioral categories. However, consistent off-diagonal structure indicates systematic sources of confusion, with both `Standing` and `Object Interaction` behaviors being misclassified as `Other` most frequently, similar to what was observed when comparing human/student annotations with model annotations on two validation videos in

Figure 6.3. This suggests errors potentially due to sensitivity to the distance threshold used to identify object interactions, as human annotations may have some tolerance to slightly smaller or larger thresholds, depending on the actual pose, and similarly sensitivity to the criterion for standing as it can imply more or less upright position.

It is also worth noting that there are no misclassifications where one specific object interaction is confused with the other. That is because the specific object the rat is interacting with is identified as the closest object/midpoint to the rat's nose. While the nose annotations may not always be very precise, they are unambiguously identifying the object given the arena's geometry.

To better understand the sources of disagreement between the YOLOv11-`large` model and the human annotations, we conducted a detailed error analysis focusing on disagreements between `Standing`↔`Other` and `Object Interaction`↔ `Other` behaviors. We observed that many discrepancies occur near behavioral transition points (i.e., when the rat transitions from one behavior to another), where subtle contextual cues are required to identify the precise frame in which the behavior changes. These transitions are often ambiguous at the frame level and depend on viewing short sequences of frames rather than isolated images. As a result, even expert annotations occasionally exhibit perceived inconsistencies in their labeling, which causes the model to make inconsistent predictions. Some examples for the `Other` ↔ `Standing` disagreements are shown in Figure 7.4 (Top Rows), while examples for the `Object Interaction` ↔ `Other` are shown in Figure 7.4 (Bottom Rows).



(a) *B06_G3*: YOLOv11-large

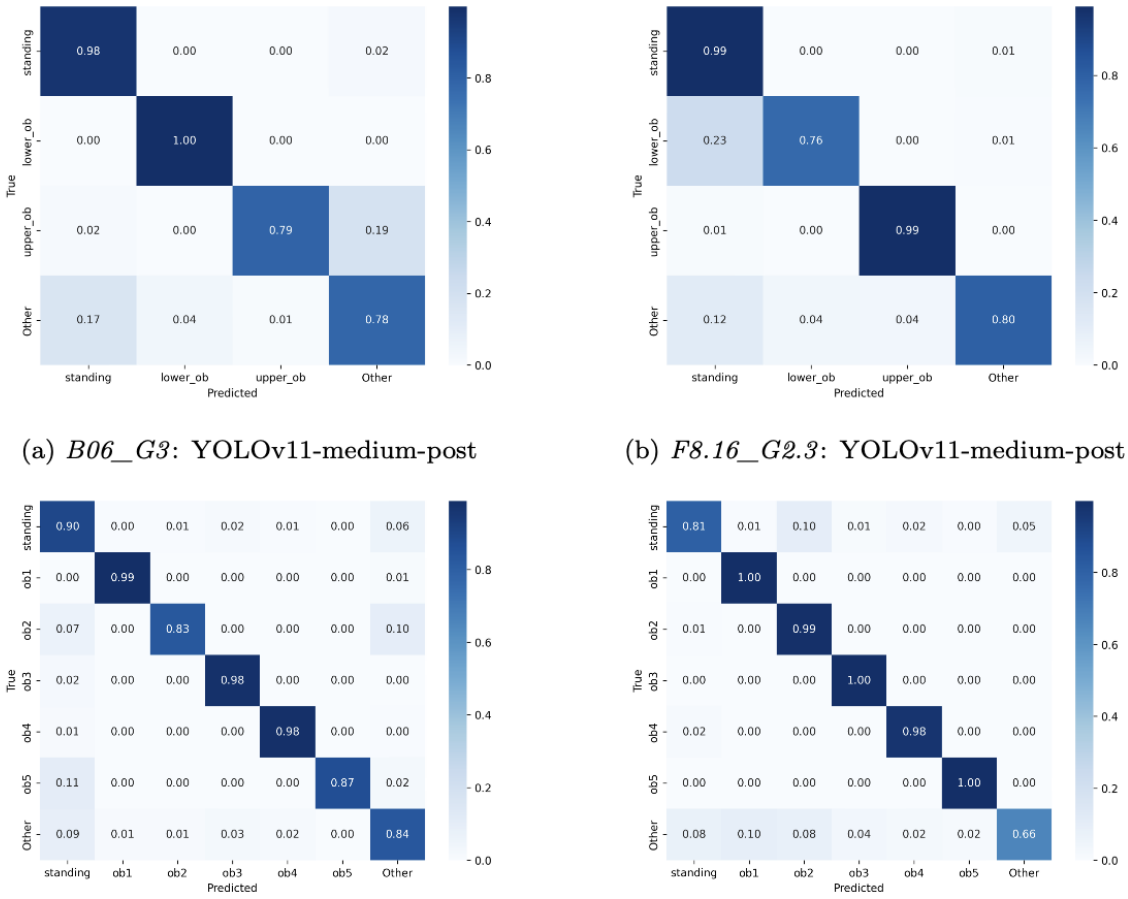
(b) *F8.16_G2.3*: YOLOv11-large

(c) *B10_B2*: YOLOv11-large

(d) *F1a.09_B3*: YOLOv11-large

Figure 7.1: *Normalized confusion matrices illustrating the classification performance of the YOLOv11-*`large` *model on the NORT behavior prediction task. Each subpanel corresponds to one test video and shows per-class agreement between model predictions and expert annotations. High diagonal values indicate strong within-class accuracy, while off-diagonal values reveal systematic confusions. Most misclassifications occur near behavioral transitions (e.g.,* `Object Interaction` ↔ `Other` *or* `Standing` ↔ `Other`*), where frame-level ambiguity and annotation inconsistencies are greatest.*

Overall, the NORT results support the use of the combined YOLOv11-`large` model, together with keypoint-based heuristic refinement, as the final behavior annotation model for downstream NORT analysis. Training on the pooled 2-object and 5-object data provided the most robust three-class predictions, while the heuristic component remained important



(a) *B06_G3*: YOLOv11-medium-post

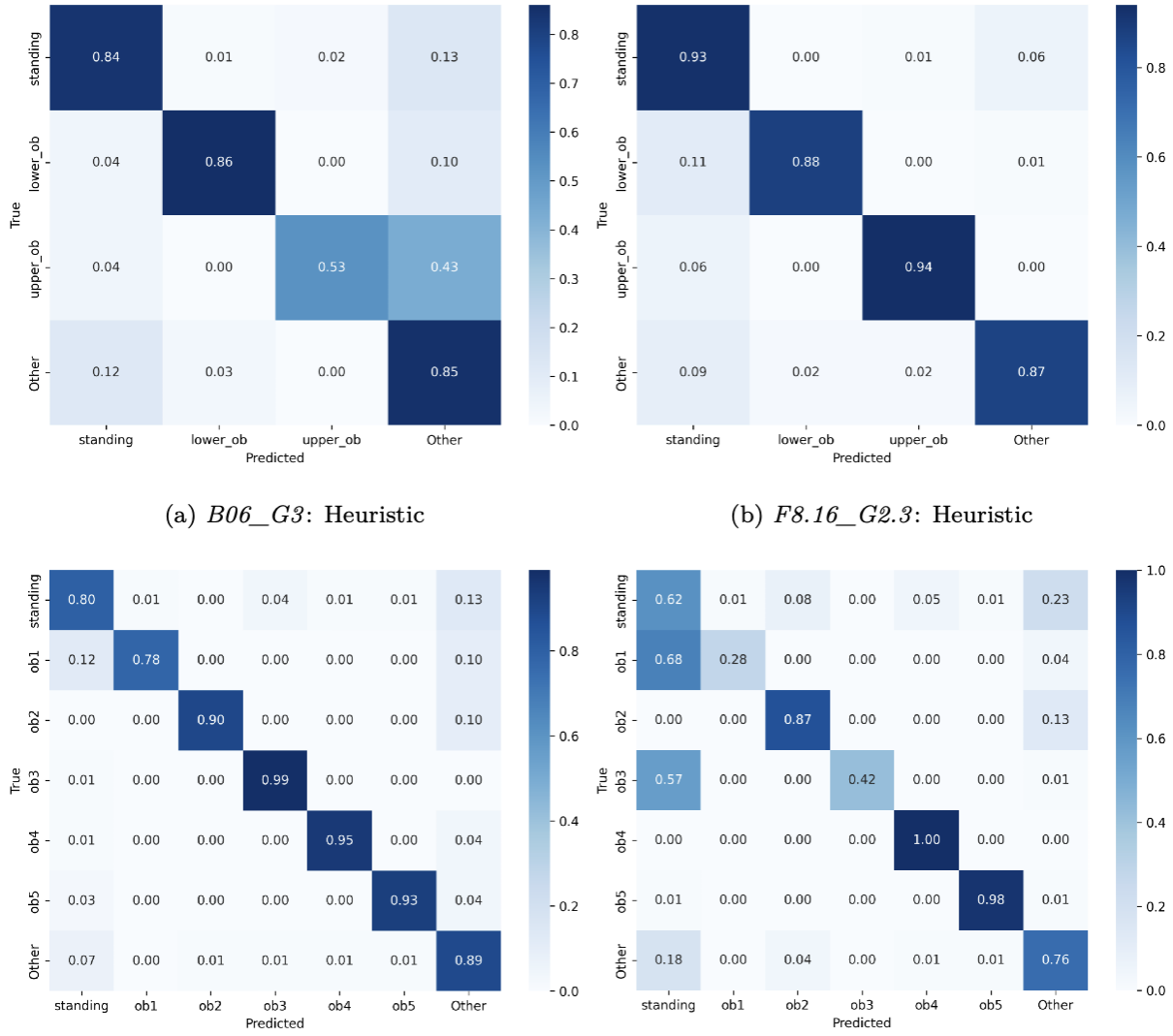
(b) *F8.16_G2.3*: YOLOv11-medium-post

(c) *B10_B2*: YOLOv11-medium-post

(d) *F1a.09_B3*: YOLOv11-medium-post

Figure 7.2: *Normalized confusion matrices illustrating the classification performance of the YOLOv11-*`medium-post` *2-object and 5-object models on the NORT behavior prediction task. Each subpanel corresponds to one test video and shows per-class agreement between model predictions and expert annotations. High diagonal values indicate strong within-class accuracy, while off-diagonal values reveal systematic confusions. Most misclassifications occur near behavioral transitions (e.g.,* `Object Interaction` ↔ `Other`*,* `Standing` ↔ `Other` *or* `Standing` ↔ `Object Interaction`*), where frame-level ambiguity and annotation inconsistencies are greatest.*

for assigning generic object-interaction predictions to specific object identities and for interpreting failure cases in relation to arena geometry. The remaining errors were concentrated primarily near behavior-transition frames, especially between Standing and Other and between Object Interaction and Other. These disagreements reflect both model limitations and the intrinsic ambiguity of frame-level behavioral annotation, where the exact onset or



(a) *B06_G3*: Heuristic

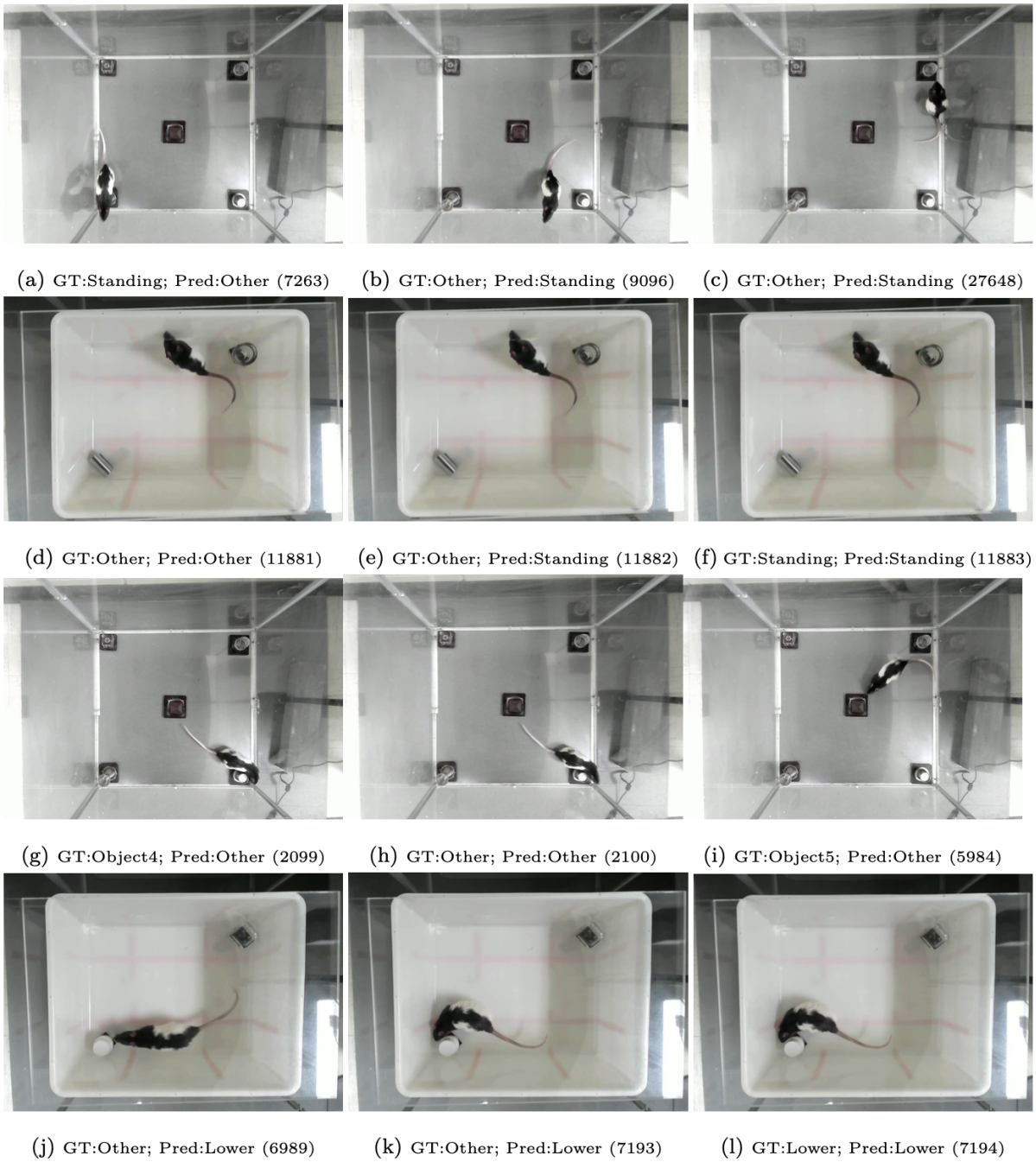
(b) *F8.16_G2.3*: Heuristic

(c) *B10_B2*: Heuristic

(d) *F1a.09_B3*: Heuristic

Figure 7.3: *Normalized confusion matrices illustrating the classification performance of the heuristic-based approach on the NORT behavior prediction task. Each subpanel corresponds to one test video and shows per-class agreement between model predictions and expert annotations. High diagonal values indicate strong within-class accuracy, while off-diagonal values reveal systematic confusions. Most misclassifications occur near behavioral transitions (e.g.,* `Object Interaction` ↔ `Other`, `Standing` ↔ `Other` *or* `Standing` ↔ `Object Interaction`*), where frame-level ambiguity and annotation inconsistencies are greatest.*

offset of a behavior may depend on subtle temporal context rather than a single isolated image.



(a) GT:Standing; Pred:Other (7263) (b) GT:Other; Pred:Standing (9096) (c) GT:Other; Pred:Standing (27648)

(d) GT:Other; Pred:Other (11881) (e) GT:Other; Pred:Standing (11882) (f) GT:Standing; Pred:Standing (11883)

(g) GT:Object4; Pred:Other (2099) (h) GT:Other; Pred:Other (2100) (i) GT:Object5; Pred:Other (5984)

(j) GT:Other; Pred:Lower (6989) (k) GT:Other; Pred:Lower (7193) (l) GT:Lower; Pred:Lower (7194)

Figure 7.4: *Error Analysis. (Top Rows) Qualitative examples for* `Standing` *vs.* `Other` *disagreements. (Bottom Rows) Qualitative examples for* `Object Interaction` *vs.* `Other` *disagreements. Each subpanel shows a single frame with the ground-truth (GT) and model prediction (Pred), along with the frame index. All frames in a row are from the same video. Some frames illustrate typical errors that may arise from frame-level ambiguity (behavior onset/offset judged over short sequences) or subtle posture cues. The consecutive frames directly illustrate the annotation challenge at the boundary between two behaviors.*

Chapter 8

Analysis-ready outputs for NORT

8.1 Final output files

After identifying the final NORT behavior annotation model, the frame-level predictions were converted into analysis-ready output files for downstream use by behavioral neuroscience collaborators. For each video, the framework generates a sequence of CSV files that preserve frame-level information while progressively adding keypoint coordinates, model predictions, object-specific interaction labels, resting/freezing labels, and final aggregated behavior labels. Table 8.1 summarizes these output files and their roles in the post-analysis pipeline.

When behavior labels overlap, `Standing` is assigned first, followed by `Object Interaction`, and then `Resting` or `Freezing` while enforcing the requirement that `Resting` and `Freezing` must be sustained for at least 60 frames. For example, if `name_4_other.csv` labels frames i through $i + 80$ as `Freezing`, but `name_3_post.csv` labels frames $i + 5$ through $i + 19$ as `Standing`, the aggregated file assigns the label `Other` to frames i to $i + 4$, as this interval does not meet the minimum threshold of 60 consecutive frames. Frames $i + 5$ to $i + 19$ are assigned the label `Standing`, while frames $i + 20$ to $i + 80$ retain the label `Freezing`. The supplementary relabeling step improves the interpretability of the generated outputs for psychological analysis workflows. Given a psychologist-defined object mapping for each 2-object video, spatial labels such as `Upper Object` and `Lower Object` are converted into

Output file	Main fields	Purpose
`name_1_keypoint.csv`	`frame#`, `nose_x`, `nose_y`, `tail_x`, `tail_y`	Stores the predicted nose and tail-base coordinates for each frame. These keypoints are used for spatial reasoning, object-assignment post-processing, resting/freezing detection, and heatmap generation.
`name_2_model.csv`	`frame#`, `behavior`	Stores the direct frame-level predictions from the YOLOv11 classification model. Behavior labels include `Standing`, `Object Interaction`, and `Other`.
`name_3_post.csv`	`frame#`, `behavior`	Combines the model predictions with keypoint-based spatial reasoning. Frames predicted as `Object Interaction` are assigned to a specific object location, such as `Upper Object` or `Lower Object` in 2-object videos, or `Object 1` through `Object 5` in 5-object videos.
`name_4_other.csv`	`frame#`, `behavior`	Uses the *nose* and *tail-base* keypoints to identify additional sustained behaviors, specifically `Resting` and `Freezing`, using the heuristic rules described in Section 5.3.
`name_5_aggregated.csv`	`frame#`, `behavior`	Merges the post-processed `Object Interaction` and `Standing` labels from `name_3_post.csv` with the `Resting`/`Freezing` labels from `name_4_other.csv`. When behavior labels overlap, priority is given to `Standing` and `Object Interaction`, with `Standing` having highest priority.
`name_6_relabeled.csv`	`frame#`, `behavior`	Uses psychologist-defined object mappings to convert spatial object labels into experimental categories such as `novel` and `familiar`. This output is primarily used for 2-object NORT videos, where object identity depends on the experimental condition.

Table 8.1: *Analysis-ready CSV files generated by the NORT post-analysis pipeline. Each file stores frame-level information, with successive files adding keypoint information, model predictions, object-specific interaction labels, resting/freezing labels, and experiment-specific relabeling.*

semantically meaningful experimental categories such as `Novel` and `Familiar`. This relabeling aligns the computational outputs with the terminology used in behavioral analysis and simplifies downstream statistical comparisons.

For distance-traveled estimates, we use the *tail-base* keypoint rather than the *nose* because the nose may move substantially during local sniffing while the body remains largely stationary. Using the *tail-base* therefore provides a more conservative estimate of whole-body locomotion and reduces inflation of distance traveled due to head movements.

8.2 Visualization outputs

In addition to these output files, the framework generates two complementary graphical visualizations derived from the files `name_1_keypoint.csv` and `name_5_aggregated.csv`. These visualizations provide researchers with both frame-level spatial tracking information and higher-level summaries of object interaction behavior, facilitating qualitative inspection alongside quantitative analysis.

8.2.1 Heatmaps

To further facilitate interpretation of the keypoint tracking data, spatial heatmaps are generated to provide a continuous representation of occupancy density throughout the experimental arena. These heatmaps aggregate positional information derived from the nose keypoints over time, thereby enabling researchers to assess the frequency and spatial distribution of visits to specific regions within the arena.

This form of visualization offers an intuitive representation of spatial patterns in addition to the discrete behavior-based summaries. This type of report captures both the intensity and spatial variability of exploration. In addition, the implementation supports grouping of experimental videos into defined cohorts, allowing for systematic comparisons between distinct sets of trial subjects. This functionality enables researchers to contrast regional patterns across populations and to identify potential deviations or irregularities in exploratory

or anxiety-related behavior.

Furthermore, the framework includes the generation of differential heatmaps, which directly visualize spatial contrasts between two groups of the same size. These difference maps highlight regions of increased or decreased occupancy relative to one another, providing a spatially resolved visualization of group-level occupancy differences and supporting more detailed downstream evaluation. This comparative approach provides a spatially resolved visualization of group-level occupancy differences and supports more detailed downstream evaluation. A generated example can be seen in Figure 8.1.

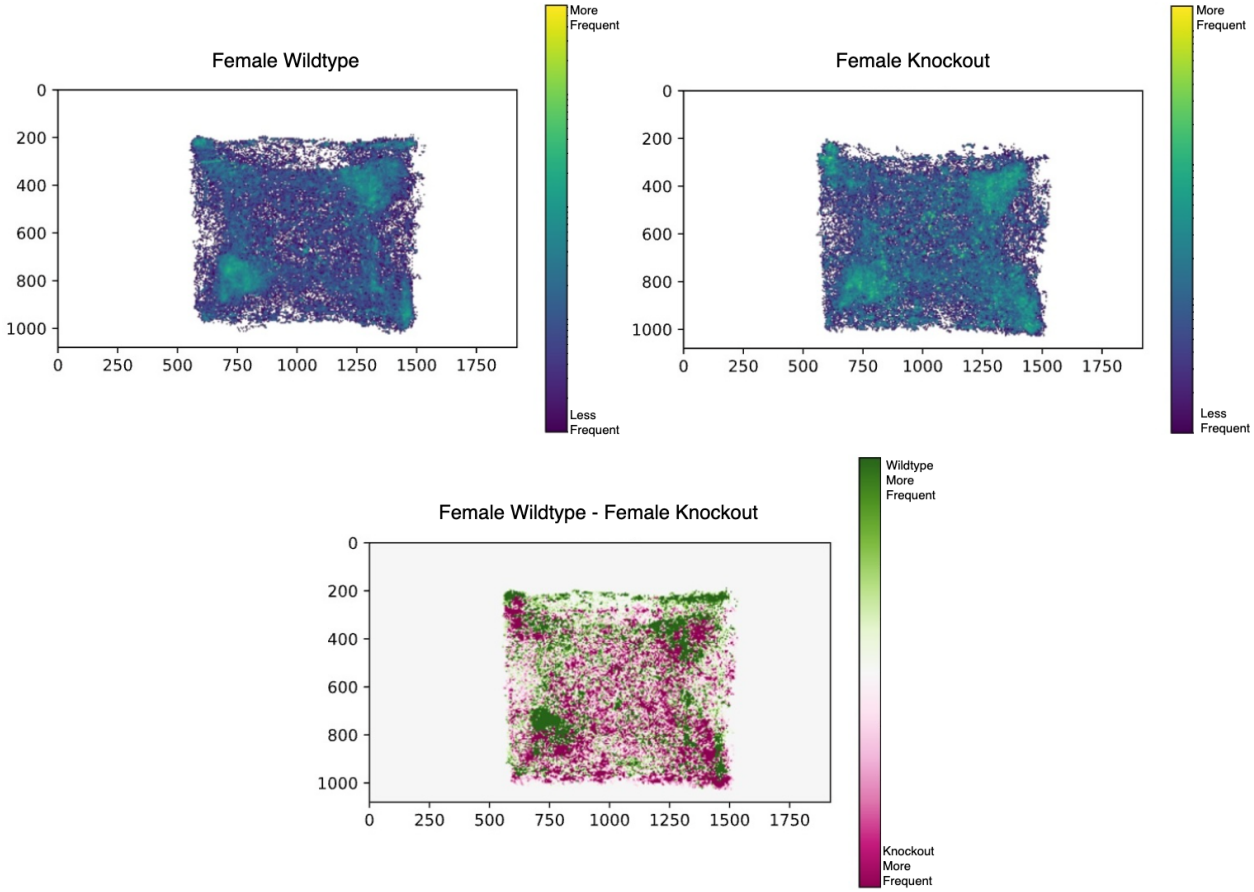


Figure 8.1: *Sample heatmaps. The top left shows the heatmap over a group of rats classified as 'wildtype'. The top right shows the heatmap over a group of rats classified as 'knockout'. The bottom shows the difference between 'wildtype' and 'knockout'. The differential map currently only supports comparison of cohorts of the same size.*

To generate this graphic, we use the process outlined in Algorithm 8.1. This algorithm requires a CSV file as input where each row has two entries; first is the name of the trial

and second is a '+' or '-' to indicate the group in which the trial belongs. Through experimentation, it was determined that a point to point heatmap was too granular, resulting in the loss of important details and excessive pixelation. To address this issue, a parameter, ϵ, was introduced to group neighboring pixels into regions rather than representing individual points. The value of $\epsilon = 3$ was preferred by our collaborators.

Algorithm 8.1 Heatmap generation algorithm.

Given *request* file ▷ user-defined, list of files for the heatmap
$\epsilon \leftarrow 3$ ▷ user-defined, how coarse/fine the heatmap will be
Given (w, h) ▷ width and height of video frames
$pluses \leftarrow zeros(0 : w, 0 : h)$
$minuses \leftarrow zeros(0 : w, 0 : h)$

$lines \leftarrow$ load *request*
for *trial, group* in *lines* **do**
    $coords \leftarrow$ load *trial*
    for $frame\#, n_x, n_y, t_x, t_y$ in *coords* **do**
        $binned_x \leftarrow floor(n_x/\epsilon) * \epsilon$
        $binned_y \leftarrow floor(n_y/\epsilon) * \epsilon$
        if $group == +$ **then**
            $pluses[binned_x, binned_y] += 1$
        else if $group == -$ **then**
            $minuses[binned_x, binned_y] += 1$
        end if
    end for
end for
plot/radiate *pluses* with user-defined color map
plot/radiate *minuses* with user-defined color map
plot/radiate $pluses - minuses$ with user-defined color map

Figure 8.2 demonstrates how the radiating of the binned (x, y) points occurs.

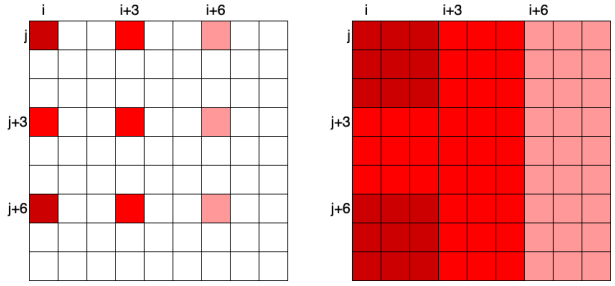


Figure 8.2: *Radiating of Binned Points.*

8.2.2 Ethograms

Supplementary to the raw behavioral output data, an additional visualization program was developed to represent the experimental results in the form of an ethogram. Ethograms are widely utilized in behavioral research as a means of illustrating the temporal progression and sequencing of behavioral states over the duration of an experimental trial. By representing behaviors along a continuous timeline, ethograms provide researchers with an intuitive and information-dense visualization of behavioral transitions, durations, and recurring patterns.

This form of representation enables complex behavioral dynamics to be interpreted in a concise and accessible manner, thereby complementing traditional numerical summaries and aggregate statistics. Rather than solely quantifying the total duration or frequency of behaviors, ethograms allow researchers to examine the temporal structure and ordering of events, which can provide additional insight into exploratory strategies, habituation patterns, and behavioral persistence.

Within the context of the NORT dataset, these visualizations can reveal nuanced behavioral tendencies that may not be immediately apparent from summary metrics alone. For example, ethograms can illustrate whether a subject consistently approaches the familiar object prior to investigating the novel object, repeatedly alternates between objects, or engages in multiple short-duration exploratory bouts directed toward the novel object. Such temporal behavioral patterns may provide meaningful insight into cognitive processes including recognition memory, exploratory drive, anxiety-related behavior, and object preference dynamics. Accordingly, the incorporation of ethogram-based visualization serves as a valuable supplementary analytical tool, enabling researchers to qualitatively and quantitatively assess the temporal organization of behavior in a manner that is both interpretable and experimentally informative.

An example of such an ethogram is shown in Figure 8.3. Each horizontal bar corresponds to an individual video trial that psychologists want to evaluate. The developed utility allows researchers to select which trials to include and in what order they are displayed. The behavior-to-color mapping is assigned via a user generated JSON file to support comparison

across trials. Additionally, behaviors can be excluded or combined to more finely examine behavior patterns. In practice, this can be used to group 'freezing' and 'resting' as 'idle' to make the visual less dense.

The ethogram-generation process is summarized in Algorithm 8.2.

Algorithm 8.2 Ethogram generation algorithm

Given request file ▷ user-defined list of trials and display order
Given behavior mapping ▷ optional behavior exclusions or behavior merges
lines ← load request file
Initialize fixed behavior-to-color map
for *trial* in *lines* **do**
    behaviors ← load `name_5_aggregated.csv` for *trial*
    Initialize *behavior_dictionary* with $\{behv_1 : [\,], behv_2 : [\,]...\}$
    for *frame_id*, *behavior* in *behaviors* **do**
        Append *frame_id* to *behavior_dictionary*[*behavior*]
    end for
    Convert each behavior list to ranges ▷ IE [2,3,4,10,11,12,13] → [(2,4),(10,13)]
    for *behavior* in *behavior_dictionary* **do**
        for $(start, end)$ in *behavior_dictionary*[*behavior*] **do**
            plot interval $(start, end)$ on the trial's horizontal bar using the behavior color
        end for
    end for
end for
Add legend, trial labels, and time-axis label

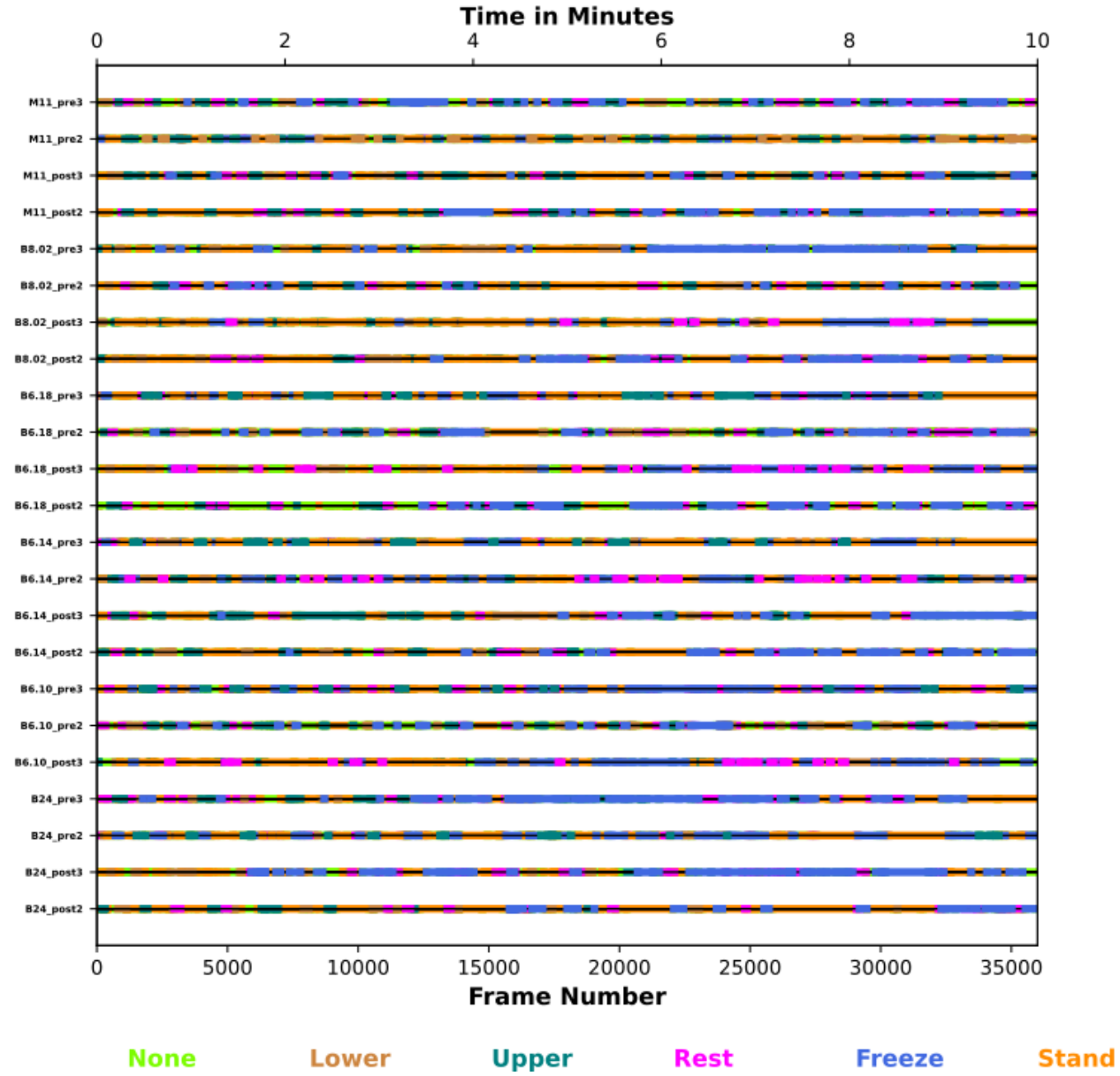


Figure 8.3: *Sample ethogram for 2-object NORT. This ethogram summarizes 23 video trials. The x-axis represents time and is shown in terms of both minutes and frames elapsed. Each color corresponds to a behavior label;* `None` *corresponds to the* `Other` *class and indicates frames in which no behavior of interest is detected.*

Chapter 9

Framework extensions for Fisher 344 NORT and EPM analysis

9.1 Overview of framework extensions

The preceding chapters developed, optimized, and evaluated the proposed framework primarily in the context of Long-Evans NORT videos. This chapter examines how the same framework can be adapted to related settings that differ either in visual appearance or in behavioral objective. First, we describe the adaptation of the NORT keypoint and behavior classification models to all-white rats, Fisher 344 rats, recorded in a similar 2-object NORT enclosure. Second, we describe the extension of the keypoint-based spatial reasoning framework to Elevated Plus Maze (EPM) videos, where the primary behavioral outputs are region occupancy and arm entries rather than object-directed interactions.

These extensions are intended to evaluate the flexibility of the proposed framework beyond the primary Long-Evans NORT setting. In particular, the Fisher 344 NORT videos provide a useful test case for visual domain adaptation because the behavioral task and enclosure remain similar, while the animal appearance differs substantially. In contrast, the EPM videos provide a test case for adapting the spatial-reasoning component of the framework to a different behavioral paradigm, where the goal is not to detect object interaction

but to summarize the animal's location across predefined regions of the apparatus.

9.2 Adaptation to Fisher 344 NORT videos

The NORT framework developed in the previous chapters was trained and evaluated primarily on videos of Long-Evans rats, which have a distinctive black-and-white coat pattern. During the course of this project, our psychology collaborators generated additional NORT videos using all-white rats, Fisher 344 rats, in a similar NORT enclosure. These videos provided an opportunity to examine whether the trained keypoint and behavior classification models could transfer directly to a visually distinct rat appearance.

Initial testing showed that the models trained on Long-Evans rats did not transfer reliably to the Fisher 344 rat videos without additional fine-tuning. The keypoint model did not consistently localize the nose and tail-base, and in some frames predicted keypoints were incorrectly placed on visually salient background structures rather than on the animal. Similarly, the behavior classification model tended to overclassify `Object Interaction` when applied directly to the Fisher 344 rat videos. These failure cases suggest that even when the behavioral paradigm and arena remain similar, differences in animal appearance and visual contrast can substantially affect both pose estimation and downstream behavior classification.

To address this domain shift and adapt the pipeline to the new visual domain, we fine-tuned both components of the NORT pipeline. First, we expanded the keypoint annotation dataset with additional manually annotated frames from Fisher 344 rat videos and used these annotations to fine-tune the YOLOv11 Pose model. Second, we expanded the frame-level behavior annotation dataset with Fisher 344 rat videos and used these annotations to fine-tune the YOLOv11 behavior classification model. The goal of this extension is not to replace the original Long-Evans NORT model, but to evaluate whether the same workflow can be adapted to a visually distinct animal appearance with a relatively modest amount of additional annotation.

9.2.1 Keypoint model adaptation

The original keypoint model was trained to detect the nose and tail-base of Long-Evans rats. When applied to Fisher 344 rat videos without additional training, the model did not reliably localize these keypoints. In particular, the reduced contrast between different body regions made it more difficult to distinguish the head, torso, and tail-base from the top-view video alone. In addition, background structures in the enclosure sometimes produced visually salient patterns that attracted incorrect keypoint predictions.

To adapt the model to this new visual domain, we manually annotated additional frames from Fisher 344 rat NORT videos. Specifically, 100 frames from two videos were annotated and divided into training and validation sets. An additional 600 frames from a third video were annotated and reserved for testing. This setup allows us to evaluate whether fine-tuning on a small number of Fisher 344 rat frames improves generalization to a held-out Fisher 344 rat video. The expanded keypoint annotation dataset is summarized in Table 9.1.

Video Type	**Video(s)**	**Train**	**Val**	**Test**	**Total**
Initial NORT	11 videos	800	200	1,200	2,200
Initial EPM	1 video	40	10	300	350
Fisher 344 NORT	Carl	40	10	0	50
Fisher 344 NORT	Luke	40	10	0	50
Fisher 344 NORT	Leslie	0	0	600	600
Total	15 videos	920	230	2,100	3,250

Table 9.1: *Expanded keypoint annotation dataset used for Fisher 344 NORT adaptation.*

After fine-tuning, the adapted keypoint model was evaluated on the 600 held-out Fisher 344 NORT test frames. The performance of the fine-tuned model can be evaluated in a spatial context, as illustrated in Figure 9.1. Across all evaluated frames, the predicted nose locations produced by the fine-tuned model were within 102mm of the corresponding ground truth positions. When extreme outliers are excluded from the analysis, this spatial error is reduced by nearly half, indicating substantially improved localization accuracy under typical operating conditions.

We compared the original Long-Evans keypoint model and the fine-tuned Fisher 344 model using the same keypoint evaluation metrics described in Chapter 4, including mAP

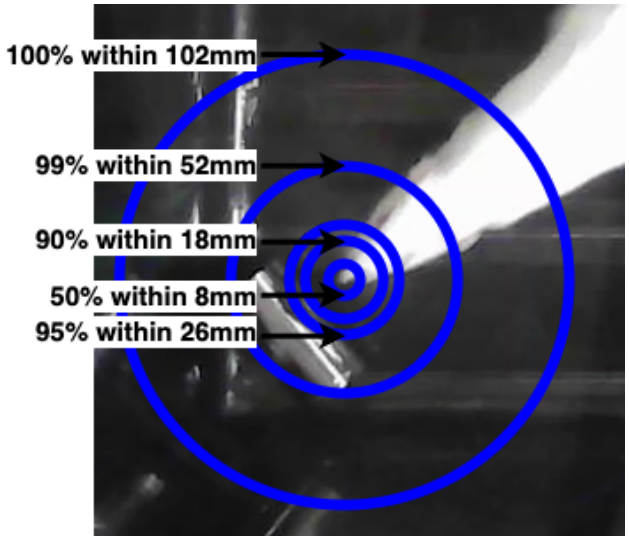


Figure 9.1: *Modified Keypoint Model on Fisher 344. Over the 600 test frames, all nose predictions are with 102mm of the ground truth, 99% of the predictions are within 52mm, and so on.*

and coordinate error for the *nose* and *tail-base*. These results are used to determine whether a small amount of additional Fisher 344 annotation is sufficient to recover reliable keypoint localization in the new visual domain. Table 9.2 reports the mAP values and mean pixel errors for both the original Long-Evans model and the fine-tuned Fisher 344 model. Of note, the original Long-Evans model only annotated 40/600 noses and tails on the test videos.

Model	Test Set	mAP@ 0.50	mAP@ 0.50:0.95	Nose Error	Tail-base Error
Original Long-Evans keypoint model	Fisher 344 NORT test	0.180	0.110	698px	593px
Fine-tuned Fisher 344 keypoint model	Fisher 344 NORT test	0.995	0.992	15px	17px

Table 9.2: *Fisher 344 keypoint model performance before and after fine-tuning. We compare performance by evaluating mAP@0.50, mAP@[0.50:0.95], and mean pixel error for the nose and tail. The original Long-Evans model annotated just* 40/600 *noses and tails while the tuned model successfully annotated noses and tails in all test frames.*

In order to evaluate and characterize the failure modes of the original Long-Evans keypoint model, a subset of erroneous frames was selected and the corresponding model predictions were superimposed onto the original images for qualitative comparison. Figure 9.2 presents a direct frame-by-frame juxtaposition of these cases, enabling visual assessment of localization accuracy across models.

The annotation scheme used in the figure is as follows: nose keypoints are indicated in green, tail keypoints in red, and the detected rat bounding box in yellow. The left column displays the predictions generated by the original Long-Evans model, while the right column shows the corresponding predictions produced by the fine-tuned model on the identical frames.

In the first two rows, the original model successfully detects the presence of the rat within the frame; however, the associated keypoint predictions exhibit substantial spatial inaccuracy. In contrast, the fine-tuned model produces markedly improved localization of both nose and tail keypoints. The final two rows illustrate more severe failure cases in which the original model misidentifies the rat entirely, whereas the fine-tuned model successfully identifies the subject and generates accurate keypoint annotations. Collectively, these examples highlight the improved robustness and detection reliability achieved through fine-tuning.

There are errors still present in the fine-tuned model which can be seen in Figure 9.3. In these errors, the model has located the rat but misplaced the nose annotation. Adding similar failure cases to the training set may improve localization accuracy and point precision in future model iterations.

9.2.2 Behavior classification adaptation

The behavior classification model also showed limited out-of-the-box transfer to the Fisher 344 videos. Although the enclosure and NORT task were similar to the original Long-Evans videos, the appearance of the animal differed substantially. Initial qualitative inspection suggested that the original behavior model tended to overclassify `Object Interaction` in the Fisher 344 videos. This behavior is problematic for downstream NORT analysis because overestimating object interaction can inflate estimates of exploration time and distort comparisons between familiar and novel objects.

To adapt the behavior model, additional Fisher 344 NORT videos were manually annotated for frame-level behavior. These annotations follow the same labeling scheme used in Chapter 5, with frames assigned to `Object Interaction`, `Standing`, or `Other`. The ex-

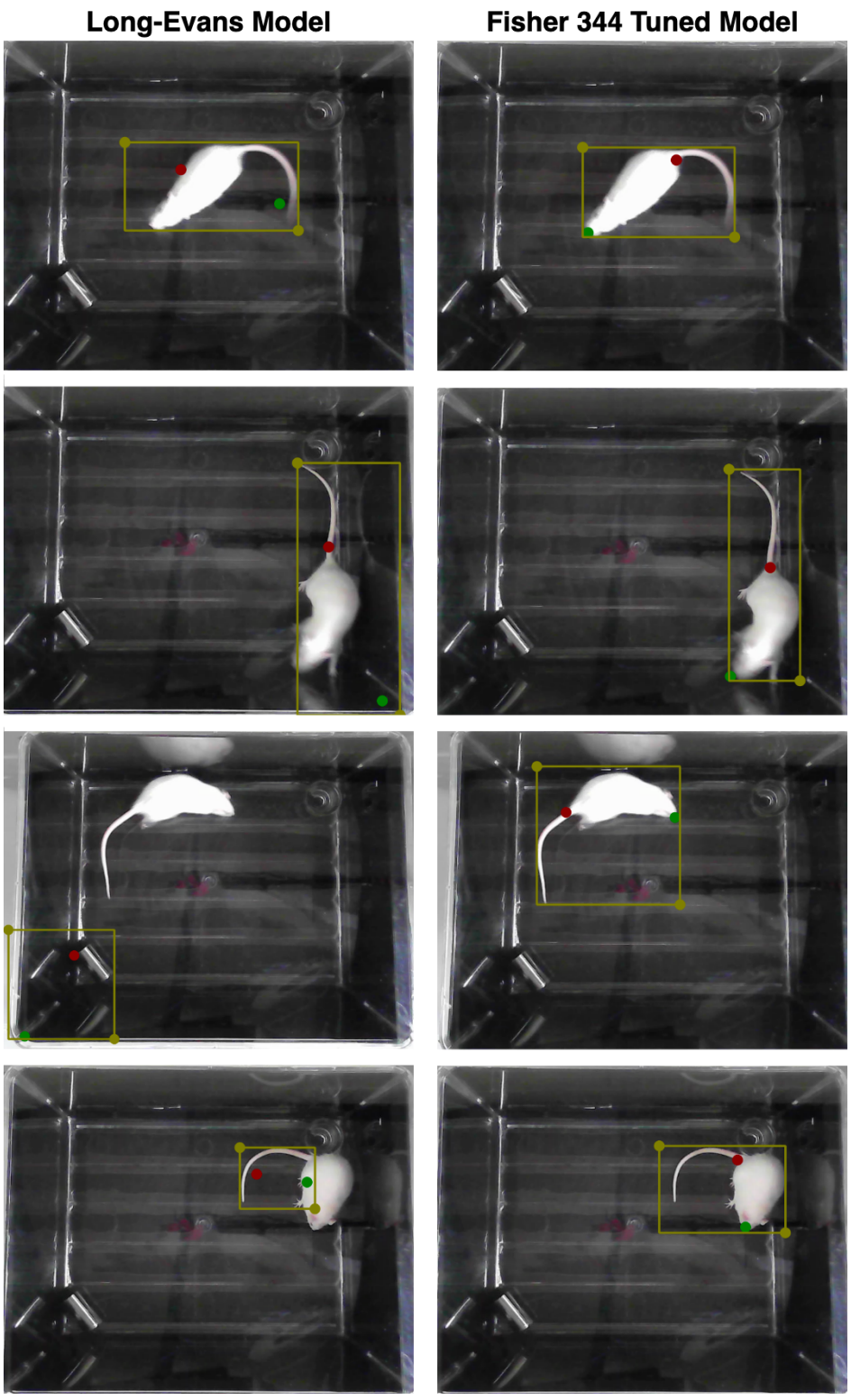


Figure 9.2: *Comparison of Long-Evans model and fine-tuned Fisher 344 models. In this figure,the nose predictions are green, the tails red and the rat bounding box yellow. The left column contains the original Long-Evans model predictions and the right column contains the fine-tuned model predictions.*

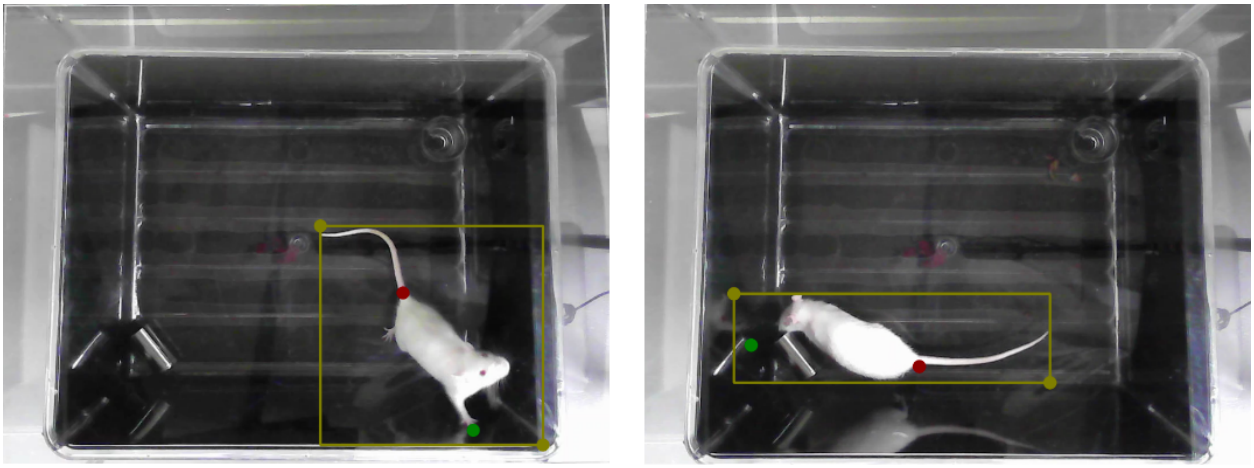


Figure 9.3: *Errors present in Fisher 344 with fine-tuned model. In this figure,the nose predictions are green, the tails red and the rat bounding box yellow. In both of these examples, the fine-tuned model is able to localize where the rat is in the frame, however, the nose annotations do not reflect the true nose position.*

panded behavior annotation dataset is summarized in Table 9.3. As with the Long-Evans rats, the Fisher 344 training and validation videos are used for fine-tuning and the held out Fisher 344 is used to evaluate the model performance.

Subset	Video(s)	Obj. Int.	Standing	Other	Total
Initial Training	3 videos	29,080	29,749	49,187	108,016
Initial Validation	3 videos	31,440	27,503	49,057	108,000
Initial Test	4 videos	29,658	41,417	72,956	144,031
Initial Data	10 videos	90,178	98,669	171,200	360,047
Fisher 344 Training	April	1,976	2,600	31,424	36,000
Fisher 344 Validation	Leslie	3,225	4,742	28,033	36,000
Fisher 344 Test	Angela	2,823	2,483	30,694	36,000
Fisher 344 Data	3 videos	8,024	9,825	90,151	108,000

Table 9.3: *Expanded behavior annotation dataset used for Fisher 344 NORT adaptation. As noted by our psychology collaborators, the Fisher 344 rats were less active than their Long-Evans counterparts.*

The fine-tuned behavior model was evaluated using the same classification metrics used for the primary NORT experiments, including precision, recall, and F1-score for each behavior class, as well as macro-F1. We compared the original Long-Evans behavior model with the fine-tuned Fisher 344 behavior model on the held-out Fisher 344 test video. To assess whether adaptation to Fisher 344 rats degraded performance on the original domain, both

models were also evaluated on a held-out Long-Evans video. The results are reported in Table 9.4.

Held Out Test Video	Model	Standing			Object Int.			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
Angela (Fisher 344)	Original Long-Evans	0.057	0.006	0.012	0.080	0.959	0.147	0.969	0.056	0.107	0.369	0.340	0.089
	Fine-Tuned Fisher 344	0.975	0.549	0.703	0.080	0.612	0.142	0.835	0.356	0.500	0.630	0.506	0.448
B06_G3 (Long-Evans)	Original Long-Evans	0.936	0.956	0.946	0.936	0.959	0.948	0.930	0.897	0.913	0.934	0.937	0.935
	Fine-Tuned Fisher 344	0.945	0.946	0.946	0.943	0.954	0.948	0.921	0.913	0.917	0.936	0.938	0.937

Table 9.4: *Behavior classification performance before and after Fisher 344 fine-tuning. Prior to fine-tuning the Long-Evans behavior model, frames from Fisher 344 videos were largely classified as* `Object Interaction`*. After fine-tuning with one video for training and one for validation, the model became more discerning about* `Object Interaction`*.*

As shown in Table 9.4, fine-tuning substantially improved overall behavior classification on the held-out Fisher 344 video, increasing macro-F1 from 0.089 to 0.448. The largest gains were observed for Standing, whose F1-score increased from 0.012 to 0.703, and Other, whose F1-score increased from 0.107 to 0.500. In contrast, Object Interaction remained challenging: its F1-score was essentially unchanged, decreasing slightly from 0.147 to 0.142. Fine-tuning reduced Object Interaction recall from 0.959 to 0.612 while precision remained 0.080, indicating that the model became less likely to assign Object Interaction indiscriminately but did not yet improve the reliability of its positive Object Interaction predictions.

Importantly, adaptation to the Fisher 344 data did not degrade performance on the held-out Long-Evans video. Macro-F1 remained essentially unchanged, increasing slightly from 0.935 to 0.937, with class-level F1-scores also remaining stable. Together, these results show that a relatively small amount of Fisher 344 annotation substantially improves recognition of Standing and Other in the new visual domain while preserving performance on the original Long-Evans domain. However, reliable Object Interaction recognition remains an important target for further adaptation.

9.2.3 Discussion of Fisher 344 adaptation

The Fisher 344 NORT extension highlights an important limitation of directly transferring computer vision models across animal appearances. Although the original and Fisher 344 videos used similar NORT enclosures and behavior definitions, the Long-Evans-trained mod-

els did not perform reliably without additional fine-tuning. This result is consistent with the broader motivation of this dissertation: automated behavioral annotation pipelines often require task- and study-specific adaptation because animal appearance, lighting, enclosure materials, and background contrast can affect both keypoint localization and behavior classification.

At the same time, the Fisher 344 extension provides a practical path forward. Rather than rebuilding the full pipeline from scratch, the existing framework can be adapted by adding a relatively small number of manually annotated frames and fine-tuning the relevant model components. The keypoint model adaptation targets visual localization failures, while the behavior model adaptation targets downstream classification errors such as overclassification of Object Interaction. Based on these experiments, approximate guidelines can be established regarding the amount of annotated data required to extend the framework to new experimental conditions. The results demonstrate that the keypoint model was comparatively straightforward to adapt to a new experimental setting, requiring only a relatively small amount of additional annotated data to achieve substantial improvements in performance. This suggests that the learned representations of the keypoint model generalize effectively across variations in experimental environment and subject characteristics.

In contrast, the behavioral classification model exhibited more moderate improvements following fine-tuning. Although the additional annotated data improved performance across several evaluation measures, the results suggest that further gains could be achieved through the incorporation of a larger set of behavioral annotations. In particular, increasing the number of annotated examples may provide the model with greater exposure to variations in posture, environmental conditions, and behavioral expression, thereby improving its ability to generalize to previously unseen experimental settings.

We further hypothesize that the behavioral classification model encountered greater difficulty adapting to the Fisher 344 dataset because of inherent differences in behavioral expression between Fisher 344 and Long-Evans rats. As noted by our psychology collaborators, Fisher 344 rats generally exhibited fewer instances of the behaviors of interest within the recorded trials and spent the majority of the time resting.

This difference in behavioral expression may partially explain the more modest improvements observed in the behavioral model following fine-tuning. In contrast to the keypoint model, which primarily learns to localize anatomical landmarks that remain relatively consistent across subjects, the behavior model must account for variation in how individual animals express and transition between behavioral states. Thus, differences in the frequency and manifestation of behaviors between rat strains may present a greater challenge for model adaptation and may necessitate a larger and more diverse set of annotated behavioral examples.

9.3 Elevated Plus Maze analysis

The Elevated Plus Maze (EPM) extension differs from the NORT setting because the primary objective is spatial region analysis rather than frame-level classification of object-directed behaviors. In NORT, the framework must determine whether the rat is interacting with an object, standing, or exhibiting no behavior of interest, and object-interaction frames must then be assigned to specific object locations. In contrast, EPM analysis focuses on determining where the rat is located within the maze over time. Therefore, the key outputs are the amount of time spent in each predefined region of the apparatus, including the open arms, closed arms, and center region, as well as the number and sequence of arm entries. This extension therefore reuses the keypoint-detection component of the proposed framework, but replaces the NORT-specific object-interaction logic with spatial region assignment and region-level behavioral summaries.

9.3.1 EPM keypoint adaptation and preprocessing

Because the EPM videos also used Long-Evans rats, the previously trained nose and tail-base keypoint model provided a strong starting point for this extension. Unlike the Fisher 344 NORT videos, the animal appearance was consistent with the primary Long-Evans NORT dataset used to develop the original keypoint model. However, the EPM setting

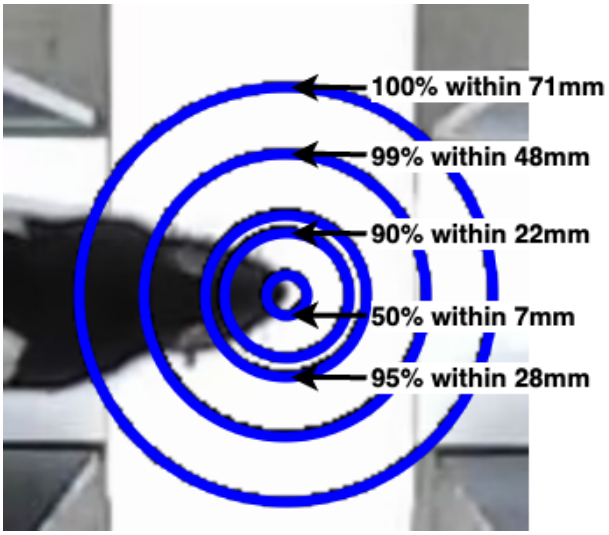


Figure 9.4: *Modified keypoint model on EPM. For the test EPM frames, each of the nose annotations were within 71mm of the actual nose position. This places the nose roughly within a head and neck length of the true nose position.*

introduced a different arena geometry and additional visibility challenges. In particular, keypoint localization became more difficult when the rat entered the enclosed arms, where the raised walls could partially obscure the body or reduce the visibility of the tail-base.

To improve robustness in these EPM-specific conditions, the keypoint model was supplemented with a small number of additional EPM annotations. Specifically, 50 targeted frames were selected from cases in which the original model had difficulty localizing the nose or tail-base, and these frames were added to the keypoint training data for fine-tuning. The performance of the keypoint model on the EPM subject can be seen in Figure 9.4. This illustrates that the nose predictions on the test frames were all within 71mm of the true nose position.

This adaptation was substantially smaller than the Fisher 344 NORT adaptation, because the rat appearance and camera perspective remained similar to the original NORT recordings. The purpose of this step was therefore not to train an entirely new model, but to adapt the existing Long-Evans keypoint model to the EPM arena layout and occlusion patterns.

Some EPM recordings also contained additional rats or visual clutter in the background, which could confound keypoint detection if the full frame were used directly. For these

videos, a fixed spatial crop was applied to isolate the EPM apparatus and target subject before keypoint inference. For cropped videos, region annotations and keypoint predictions were generated within the same cropped coordinate system, ensuring that the predicted nose and tail-base coordinates could be used directly for subsequent region assignment. Together, the supplemental annotations and cropping procedure provided the keypoint inputs needed for frame-level EPM spatial analysis.

9.3.2 EPM region definition

After keypoint prediction, the next step in the EPM pipeline is to define the spatial regions of the maze. Unlike NORT, where the relevant spatial regions correspond to objects in the arena, EPM analysis requires partitioning the apparatus into the north arm, south arm, east arm, west arm, and center region. These regions are then used to convert frame-level keypoint predictions into frame-level location labels.

For each trial, the EPM apparatus is given an initial annotation by the user which defines the square where the arms intersect. Given the center intersection the region boundaries are mathematically interpolated. These annotations identify the major geometric structure of the maze and provide the basis for computing consistent region boundaries across videos. In the videos analyzed here, the north and south arms are the open arms, while the east and west arms are the enclosed arms. Figure 9.5 illustrates the EPM region-definition procedure. The solid blue points and lines indicate user-defined reference annotations, while the dashed black lines and black points indicate interpolated boundaries used to partition the maze into arm and center regions. This procedure provides a consistent geometric representation of the EPM arena for downstream frame-level region assignment.

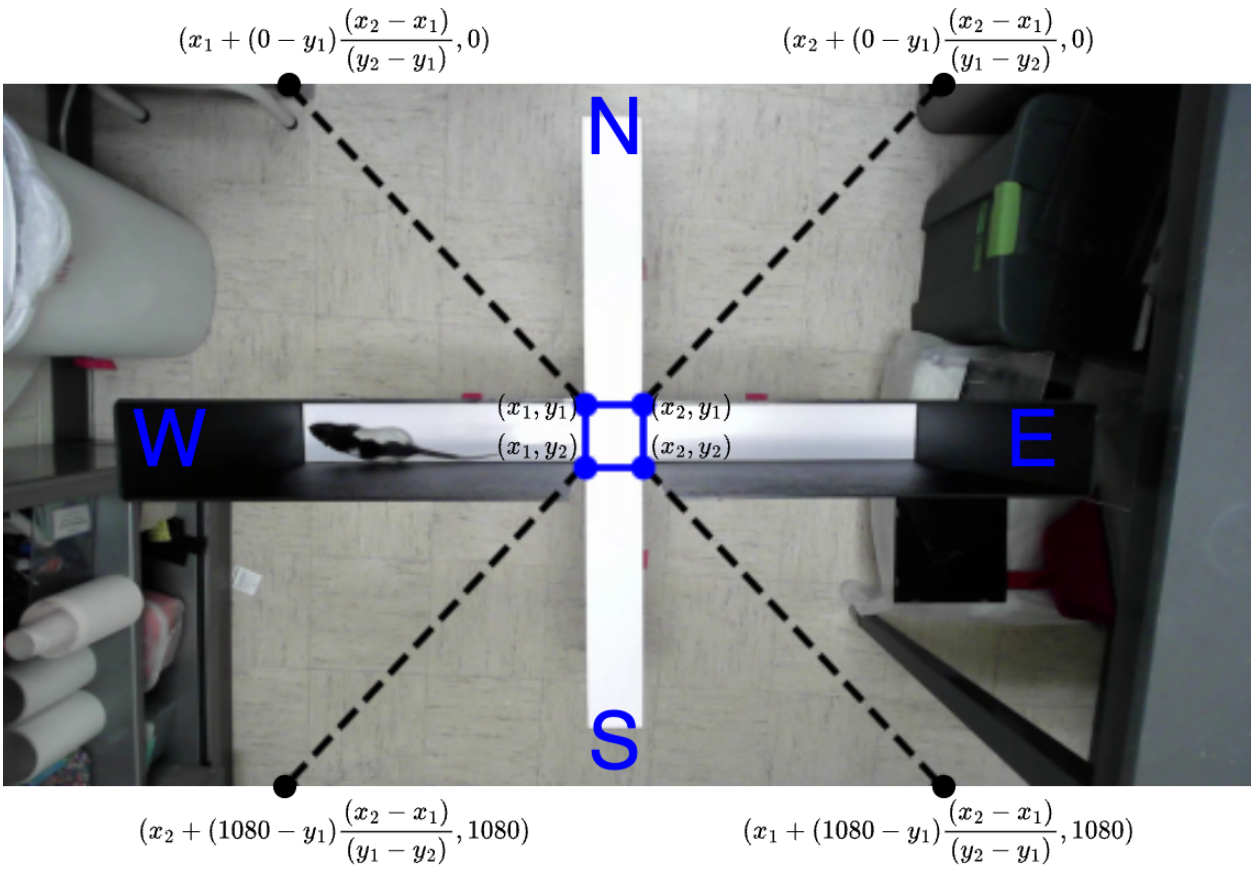


Figure 9.5: *EPM region definition procedure. Solid blue points and lines indicate user-defined reference annotations, while dashed black lines and black points indicate interpolated boundaries used to partition the maze into arm and center regions. These five regions are used to assign frame-level keypoint predictions to EPM spatial labels.*

9.3.3 Frame-level region assignment and bout counting

Once the EPM regions are defined, the predicted nose and tail-base keypoints are used to assign each video frame to a spatial region. A frame is assigned to a given EPM region only when both the nose and tail-base keypoints fall within the boundaries of that region. This conservative criterion is intended to capture full-body occupancy rather than partial or transitional positioning. If the nose and tail-base fall in different regions, or if either keypoint lies in an ambiguous boundary area, the frame is assigned to an `Other` category.

This full-body criterion reduces false region assignments during transitions between arms and the center region. However, it also means that some frames corresponding to partial

entries, boundary crossings, or risk-assessment postures may be assigned to `Other`. This tradeoff is appropriate for the current EPM analysis because the primary goal is to obtain robust estimates of time spent in each region and the sequence of region occupancy episodes rather than to classify more subtle ethological behaviors such as head dips or stretch-attend postures.

The frame-level region labels are then converted into occupancy bouts. A bout is defined as a maximal contiguous sequence of frames assigned to the same EPM region. For example, if a rat remains in the north open arm for a continuous sequence of frames, that sequence is counted as one north-arm bout. A minimum bout duration is applied to remove short, noisy transitions; in the current implementation, this threshold is set to 15 frames, corresponding to a quarter of a second. Arm entries are derived from these bouts by counting transitions into one of the arm regions from the center or `Other` category. Open-arm entries and closed-arm entries can then be summarized separately by grouping the north and south arms as open arms and the east and west arms as closed arms.

9.3.4 EPM summary outputs and visualizations

The EPM pipeline produces both frame-level and summary-level outputs. At the frame level, the framework stores the predicted nose and tail-base coordinates and the corresponding region label for each frame. At the summary level, the framework computes the total time spent in each region, the number of entries into each arm, the number and duration of occupancy bouts, and aggregate open-arm versus closed-arm measures. These outputs provide the main behavioral summaries required for downstream EPM analysis.

In addition to numerical summaries, the EPM extension supports spatial and temporal visualizations similar to those generated for NORT. Heatmaps can be used to visualize where the rat spent the most time within the EPM apparatus, as shown in Figure 9.6, while ethograms can be used to show the temporal sequence of region occupancy across a trial, as shown in Figure 9.7.

These visualizations complement aggregate measures by showing the order of region vis-

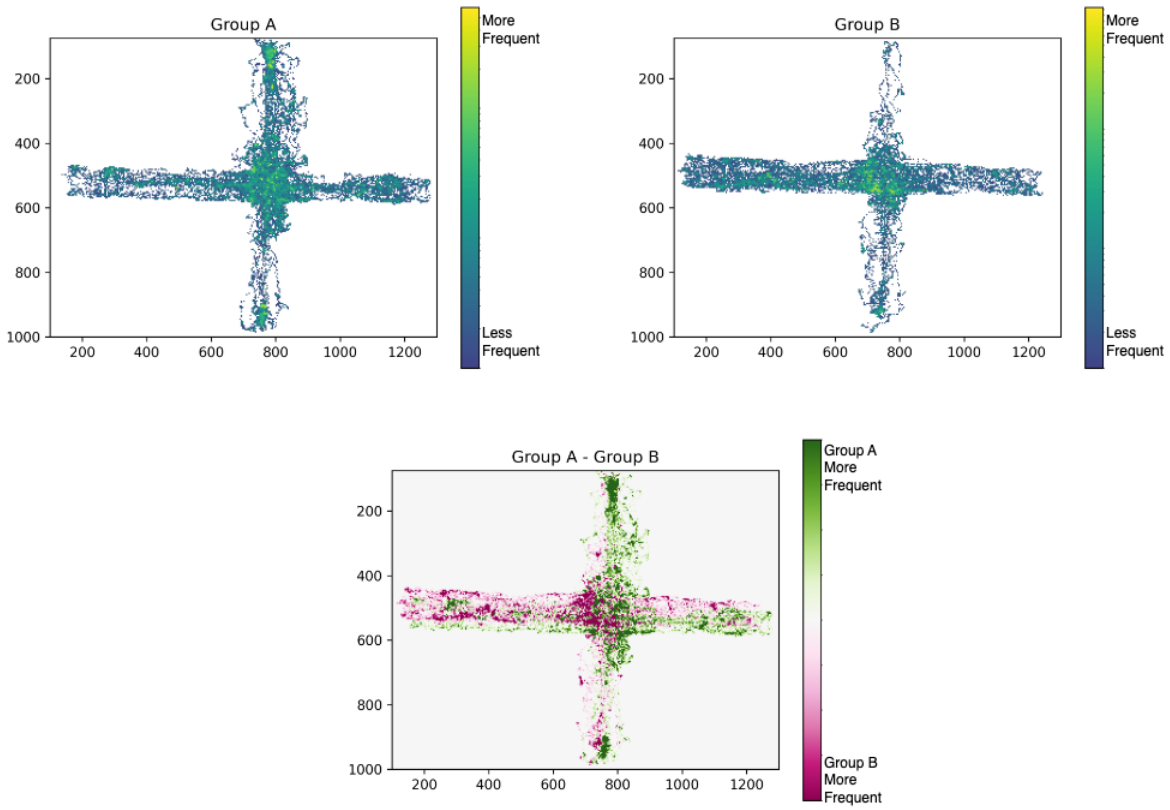


Figure 9.6: *EPM heatmap. Similar to the output options available for the NORT dataset, spatial heatmaps can also be generated for EPM trials. In this rendering, each experimental group consisted of four subjects*

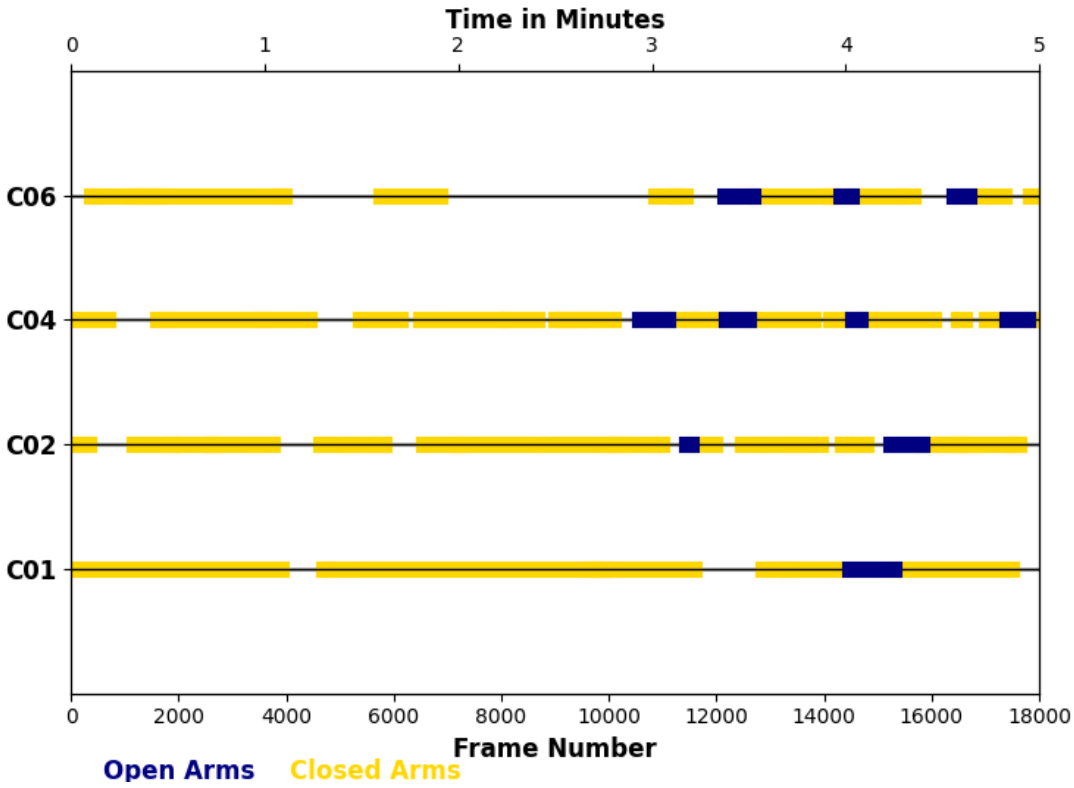


Figure 9.7: *EPM ethogram. For EPM trials, ethograms can be utilized to capture patterns of region exploration. In these particular subjects, the open arms were not explored until the second half of the trial.*

its, the duration of individual occupancy bouts, and spatial preferences within the open and enclosed arms. The final EPM outputs therefore adapt the same general framework used for NORT to a different behavioral objective. Rather than identifying object-directed behaviors, the EPM extension uses keypoint predictions and region definitions to generate spatial occupancy summaries, arm-entry counts, bout-level summaries, heatmaps, and ethograms. This demonstrates how the keypoint-based spatial reasoning component of the proposed framework can be reused across behavioral paradigms with different analysis goals.

Chapter 10

Marble burying and marble-interaction analysis

This chapter extends the automated behavioral annotation framework developed for NORT and adapted to EPM to a third behavioral paradigm: marble-interaction analysis in the Marble Burying Test (MBT). Unlike the NORT setting, where objects are fixed in known locations, the MBT setting contains multiple small objects/marbles that may move, become partially occluded, or become buried over the course of the trial. Therefore, this task requires both rat keypoint tracking and marble localization/tracking before frame-level behavioral labels can be inferred.

The MBT results in this chapter should be interpreted as a preliminary extension of the proposed framework rather than a fully optimized MBT behavior-recognition system.

10.1 Overview and behavioral objectives

The additional dataset explored in this chapter corresponds to the Marble Burying Test (MBT). In this experimental setup, a single rat is placed within a plastic enclosure containing a layer of wood-fiber bedding material designed to permit natural digging and exploratory behaviors. At the onset of each trial, an array of twenty marbles is positioned evenly across

the surface of the bedding, providing a standardized arrangement for behavioral assessment.

The experimental recordings consist of overhead video captured from an aerial, or "eagle-eye," perspective, enabling observation of the animal's movement and interactions throughout the enclosure. Videos were recorded at 15 frames per second with a resolution of (480×480) pixels. Although the video quality is lower than that of the NORT videos, the setup provides sufficient temporal continuity and spatial detail to support automated analysis of rat position, marble location, and frame-level marble-directed behavior.

The traditional MBT outcome is the number or percentage of marbles buried at the end of the trial. While this endpoint measure is useful, it provides only a coarse summary of behavior. It does not capture the temporal progression of the trial, the sequence of interactions that led to marble displacement, or the distinction between deliberate marble-directed behaviors and incidental substrate movement caused by ordinary locomotion. As a result, endpoint scoring alone may obscure important behavioral dynamics occurring during the trial.

The objective of the analysis presented in this chapter is to move beyond endpoint marble counts by extracting frame-level behavioral information from the videos. In particular, the current analysis focuses on five mutually exclusive frame-level labels: `Marble Interaction`, `Standing`, `Resting`, `Freezing`, and `Other`. The Marble Interaction label captures frames in which the rat is positioned near and appears behaviorally directed toward a marble. The Standing label captures rearing or upright postural behavior, including cases where the rat stands against the arena wall. The resting and freezing labels correspond to idle behaviors as described in the NORT trials. The Other label includes all remaining frames, including locomotion, non-marble-directed exploration, and behaviors not separately annotated in the current dataset.

By providing frame-level labels and marble trajectories, the proposed MBT extension gives neuroscientists a richer representation of trial dynamics than traditional endpoint scoring. In particular, it supports analysis of when marble-directed behavior occurs, how long such behavior persists, which marbles are contacted or displaced, and how marble positions change over time. This information may help distinguish marbles that appear to have been

intentionally displaced or buried through directed interaction from marbles that become partially covered as an indirect consequence of the rat traversing or disturbing the bedding material.

Although the broader MBT pipeline can support `Marble Interaction`, `Standing`, `Resting`, `Freezing`, and `Other` labels, the supervised YOLOv11 behavior-classification experiments in this chapter focus on the three manually annotated classes: `Marble Interaction`, `Standing`, and `Other`. `Resting` and `Freezing` are handled using heuristics and are thus not included in the MBT classifier evaluation.

10.2 Rat keypoint adaptation for MBT videos

As with the previously described experimental paradigms, the MBT analysis requires accurate localization of rat body keypoints. In this work, the keypoint model predicts the *nose* and *tail-base* of the rat. These two keypoints provide a compact representation of the animal's position, orientation, and posture, and they serve as the foundation for downstream behavioral analysis. In particular, the *nose* keypoint is used to estimate proximity to nearby marbles, while the vector from the *tail-base* to the *nose* provides information about the direction in which the rat is oriented.

The keypoint predictions also support additional behavioral measures beyond marble interaction. For example, changes in the distance between the *nose* and *tail-base* can be used to detect postural changes associated with `Standing`. In addition, continuous tracking of keypoints enables computation of total distance traveled, movement speed, and low-mobility periods that support `Resting` or `Freezing` analysis.

To support robust performance within the MBT setup, the keypoint dataset described earlier in Table 9.1 was expanded to include additional annotated examples from the marble trial recordings. The resulting dataset is summarized in Table 10.1. These additional frames expose the model to the visual conditions of the MBT environment, including the bedding texture, lower video resolution, and rat postures that occur during digging and marble-directed exploration.

Video Type	Video(s)	Train	Val	Test	Total
NORT (Long-Evans)	11 videos	800	200	1,200	2,200
EPM	1 video	40	10	300	350
NORT (Fisher 344)	3 videos	80	20	600	700
Marble	48.06	90	10	0	100
	78.03	0	0	611	611
Total	17 videos	1,010	240	2,711	3,961

Table 10.1: *Expanded keypoint annotation dataset used for MBT keypoint-model adaptation. The marble-trial frames were added to improve model robustness under the visual conditions of the MBT apparatus. In comparison to the original dataset, a small amount of annotations were required for fine-tuning.*

Only a modest amount of additional annotation was required to adapt the existing keypoint model to the marble-trial setting. Specifically, 100 frames from a single marble video were manually annotated and used for training. A different video was utilized exclusively for evaluation, with 611 frames sampled at one frame per second to assess the model's performance. This suggests that the keypoint model trained on the earlier NORT and EPM data retained substantial transferable representational capacity and could be adapted efficiently to the MBT videos.

Figure 10.1 shows an example of the nose keypoint verification for the adapted MBT keypoint model. The maximum distance between the ground truth and prediction was 104mm or 66 pixels. The predicted nose location is used directly in the marble-interaction heuristic described in Section 10.4; therefore, accurate nose localization is particularly important for this task.

To further analyze the improvements offered by fine-tuning to the marble environment, mAP@0.50 and mAP@[0.50:0.95] were evaluated, as well as the mean pixel error. These metrics are outlined in Section 4.3. When utilizing mAP, values closer to one reflect higher accuracy. The mean pixel error was also measured independently for the nose and tail. The results of these metrics can be seen in Table 10.2. A notable observation is the disparity in tail keypoint detection performance between the two models. The initial Long-Evans keypoint model successfully identified all nose keypoints 611/611; however, it correctly detected only 494/611 tail keypoints. In contrast, the fine-tuned MBT keypoint model maintained perfect

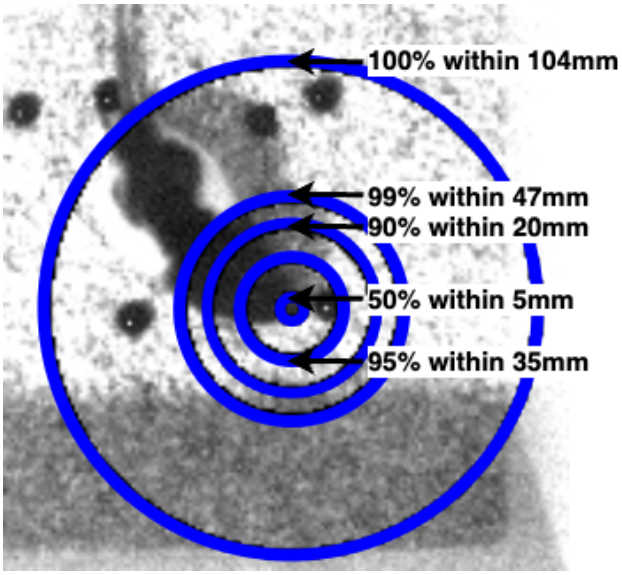


Figure 10.1: *Example nose keypoint verification for the MBT keypoint model. Across the 611 testing frames, the nose prediction was within 104mm of the true position. The predicted nose location is evaluated because it is the primary keypoint used for estimating proximity to marbles during downstream marble-interaction analysis.*

nose keypoint detection 611/611 while substantially improving tail keypoint localization, successfully identifying 609/611 tail keypoints. This improvement suggests that the fine-tuning process significantly enhanced the model's ability to localize the more challenging tail keypoint, while preserving the already high accuracy of nose detection.

Model	Test Set	mAP@ 0.50	mAP@ 0.50:0.95	Nose Error	Tail-base Error
Original Long-Evans keypoint model	MBT test	0.513	0.337	167px	232px
Fine-tuned MBT keypoint model	MBT test	0.990	0.967	5px	5px

Table 10.2: *MBT keypoint model performance before and after fine-tuning. Comparison of mAP@0.50, mAP@[0.50:0.95], and mean pixel error on the held out 611 evaluation frames showed substantial improvements in model performance. Both models successfully detected a nose keypoint in every evaluation frame. However, the original model identified only 494 tail keypoints, whereas the fine-tuned model predicted 609, representing a substantial improvement in tail localization performance.*

To further illustrate the improvements after fine-tuning, Figure 10.2 was generated where the nose keypoints are red, the tail keypoints are green and the rat bounding box is yellow. The left column illustrates representative failure cases produced by the original Long-Evans keypoint model, while the corresponding frames in the right column show the predictions

generated by the fine-tuned model. Each row presents the same video frame, allowing for a direct qualitative comparison between the two models.

In the first row, the original model incorrectly identifies the subject rat as a handle located in the upper-left corner of the experimental arena, resulting in a failed keypoint prediction. In contrast, the fine-tuned model correctly detects the rat and accurately localizes the nose and tail keypoints. The second row depicts a frame in which the original model incorrectly predicts the nose keypoint near the base of the tail while failing to detect the tail keypoint altogether. The fine-tuned model, however, correctly identifies both keypoints with high spatial accuracy. The final two rows present additional examples in which the original model predicts both the nose and tail keypoints but localizes them incorrectly, compressing the annotations toward the center of the rat's body and placing them a considerable distance from their true anatomical positions. By comparison, the fine-tuned model produces keypoint predictions that closely correspond to the ground truth locations.

Collectively, these qualitative examples complement the quantitative evaluation presented earlier, demonstrating that fine-tuning substantially improves both the robustness of subject detection and the accuracy of nose and tail keypoint localization across a variety of challenging frames. While there are distinct and notable improvements, minor errors can be seen in select frames. Examples of these errors can be seen in Figure 10.3. Identifying these errors can guide further improvements to the keypoint model by incorporating the corresponding frames into the training dataset.

10.3 Marble segmentation and identity tracking

In addition to rat keypoints, the MBT analysis requires accurate localization of the marbles. This requirement differs from the NORT object-interaction setting. In NORT, objects are fixed at known locations throughout the trial, and therefore object regions can be defined manually as static bounding boxes or circular regions. In MBT, however, marbles may be displaced over time as a result of rat movement, digging, or direct marble interaction. Consequently, the locations of the marbles must be estimated dynamically.

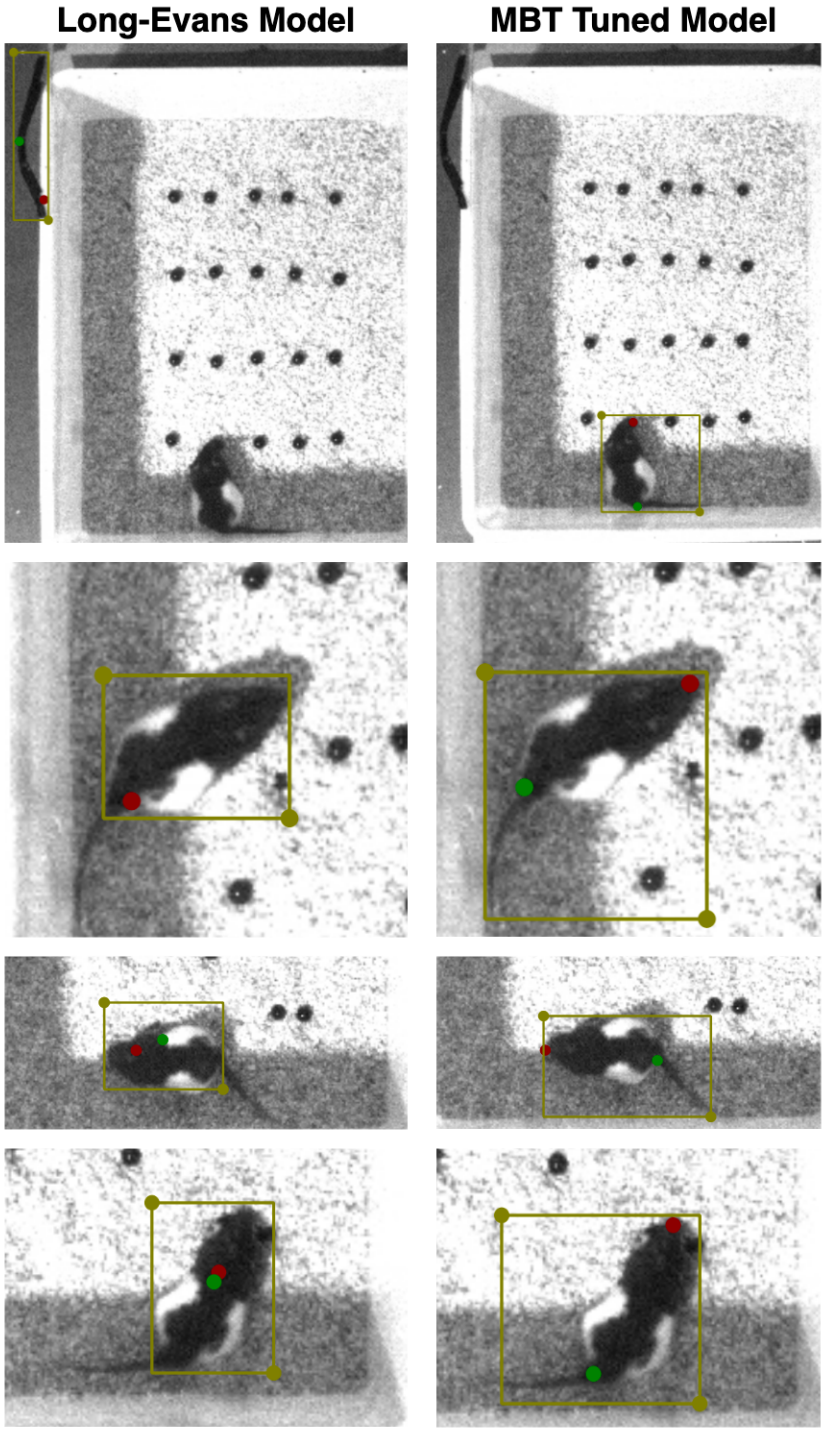


Figure 10.2: *Comparison of Long-Evans model and fine-tuned MBT models. In this figure,the nose predictions are red, the tails green and the rat bounding box yellow. The left column illustrates examples of incorrect annotations done by the original Long-Evans model and the right column shows the corresponding improvement offered by the fine-tuned model.*

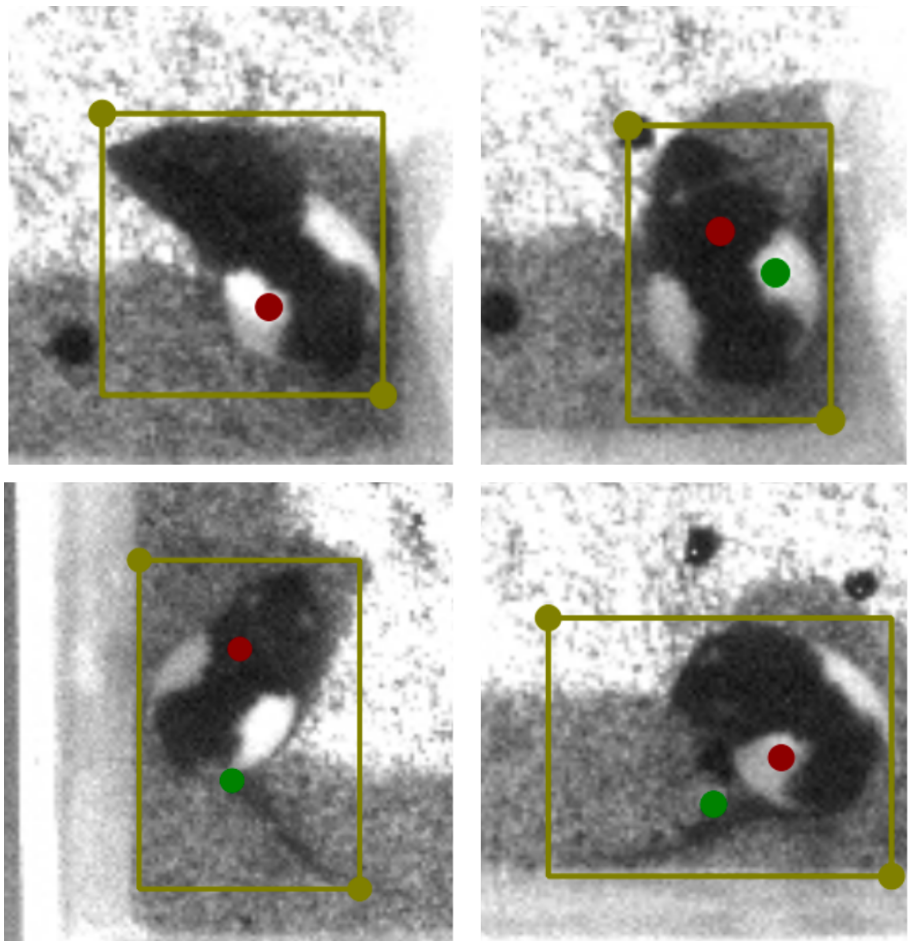


Figure 10.3: *Errors present in MBT with fine-tuned model. In this figure,the nose predictions are red, the tails green and the rat bounding box yellow. The top left frame is an example of a tail keypoint missing. The remaining images show examples where the rat has been localized but the keypoints do not match the true anatomical positions.*

Explicit marble localization provides two important benefits. First, it enables frame-level interaction detection by allowing the system to determine whether the rat's nose is sufficiently close to a marble. Second, it allows the system to estimate marble displacement over time, which may provide useful information about the temporal development of burying behavior. Because marbles can become partially or fully obscured by bedding, persistent tracking also helps maintain an estimate of marble identity and location when a marble is temporarily difficult to detect.

Consistent with the computer-vision workflow used in the rest of this dissertation, a segmentation-based YOLOv11 model was used to detect marble instances in individual frames. Specifically, the YOLOv11-medium segmentation model was fine-tuned. For each predicted marble region, the center point was computed and used as the estimated marble position. Given an estimated or predefined marble radius, this center point can be used to reconstruct the approximate spatial extent of the marble, including cases in which the visible portion of the marble is partially obscured by bedding.

The dataset used to train and validate the marble segmentation model is summarized in Table 10.3. In line with the bootstrapping method outlined in Section 4.2, initial training was done on 220 frames from 2 videos and predicted frames were visually inspected to identify incorrect predictions. The resulting dataset including 5 videos for training and validation with 340 frames. This includes frames in which some marbles were partially covered or buried. Across these frames, 6,166 marble instances were annotated. For evaluation of the model, 100 frames were selected from a a separate video with a total of 1,901 marbles annotated. Annotation of the marble segments was done using the LabelMe interface Russell et al. (2008).

An illustrative example of marble center prediction is shown in Figure 10.4 where marble centers are annotated in red. The displayed frame includes the frame number and the distance from the predicted nose keypoint to the nearest detected marble center. In this example, the frame corresponds to the third frame of the video, and the rat's nose is approximately 15.03 pixels from the nearest marble.

Because marble localization is formulated as an instance segmentation task, the segmen-

Video	Annotation Type	Train	Val	Test	Total
47.06	Frames	89	22	0	111
	Marble instances	1,742	434	0	2,176
47.07	Frames	100	10	0	110
	Marble instances	1.917	194	0	2,111
48.06	Frames	3	0	0	3
	Marble instances	57	0	0	57
49.06	Frames	4	2	0	6
	Marble instances	80	40	0	120
50.05	Frames	100	10	0	110
	Marble instances	1,514	188	0	1,702
78.03	Frames	0	0	100	100
	Marble instances	0	0	1,901	1,901
Total	Frames	296	44	100	440
	Marble instances	5,310	856	1,901	8,067

Table 10.3: *Marble segmentation dataset used for marble localization. The annotated frames include cases in which some marbles were partially covered by bedding, allowing the model to learn from visually challenging examples. Across the training, validation, and test subsets, a total of 440 frames containing 8,067 marble instances were annotated.*

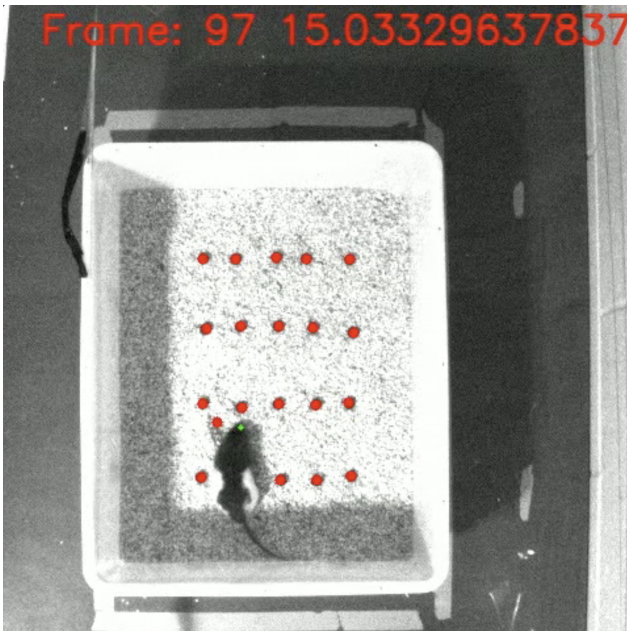


Figure 10.4: *Example marble predictions from the marble segmentation model. The red points indicate the marble centers with a radius of 4 pixels. The displayed annotation includes the frame number, 97, and the distance from the rat's nose to the nearest detected marble. In this example, the nose, in green, is approximately 15.03 pixels from the nearest marble.*

tation model was evaluated using standard object detection metrics, including precision, recall, mAP@0.50, and mAP@[0.50:0.95]. These metrics provide complementary measures of detection performance, with precision quantifying the proportion of predicted marbles that correspond to true marble instances, recall measuring the proportion of ground truth marbles successfully detected, and the mean Average Precision (mAP) metrics evaluating localization accuracy across one or more Intersection over Union (IoU) thresholds.

The segmentation model was evaluated on a test set containing 1,901 annotated marble instances. The model obtained an mAP@0.50 of 0.438, demonstrating moderate segmentation accuracy at the standard IoU threshold of 0.50. However, performance declined substantially under more stringent evaluation criteria, with an mAP@[0.50:0.95] of 0.0972. This decrease suggests that while the model was generally able to localize marble instances, it had difficulty producing consistently accurate segmentation masks at higher IoU thresholds. Despite the relatively low mAP values, these results are not prohibitive for the downstream application. The subsequent stages of the behavioral analysis pipeline rely solely on the centroid of each detected marble rather than the accuracy of the complete segmentation mask. Therefore, accurate localization of the marble center is of greater practical importance than achieving high segmentation overlap, making centroid detection a more relevant measure of performance for this task.

Further improvements in segmentation performance would be expected through the inclusion of a larger and more diverse annotated training dataset. However, generating high-quality segmentation annotations for individual marbles is a labor-intensive and time-consuming process, particularly due to the frequent partial occlusion of marbles by bedding material and the requirement for pixel-level mask annotations. Consequently, expanding the training dataset represents a significant manual annotation effort despite its potential to substantially improve model performance.

Beyond frame-level marble detection, it is useful to maintain consistent identities for individual marbles across the trial. Persistent marble identities allow downstream analysis to quantify how far each marble moves, when displacement occurs, and whether particular marbles are repeatedly contacted by the rat. To support this, each marble is assigned an

identity at the beginning of the trial, and detections in later frames are matched to previously tracked marble states. Figure 10.5 illustrates the initial configuration of the marbles on the left and a subsequent frame from the same trial on the right, demonstrating the displacement of the marbles over the course of the experiment.

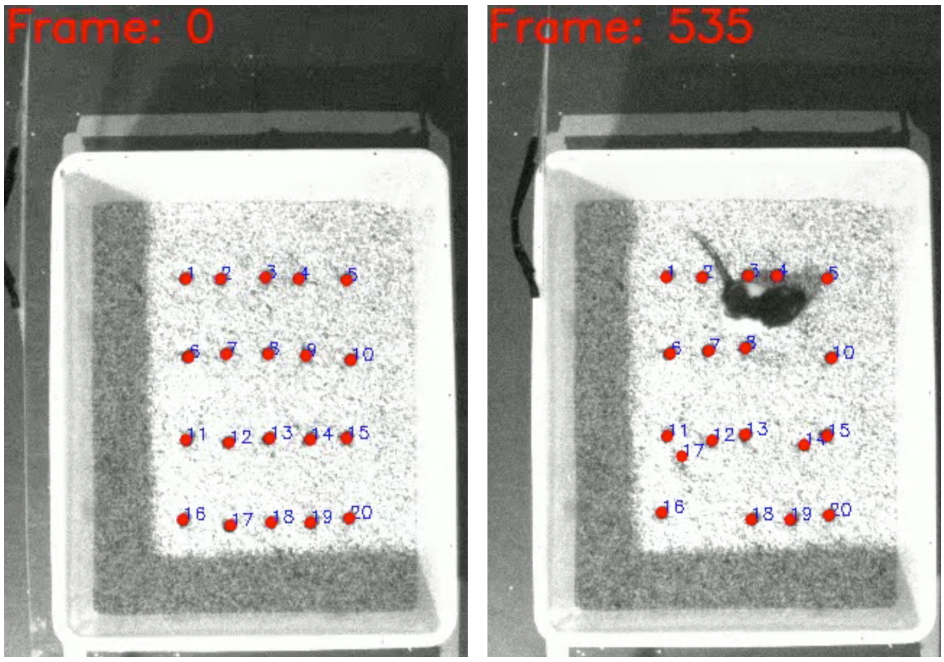


Figure 10.5: *Marble identification. The original configuration of the trial with marbles identified numerically is shown on the left. Frame 535 is shown on the left with the persistent numeric labels. In this frame, the rat is covering and investigating marble 9.*

Throughout the tracking of the marbles, a last known location is stored for each marble. If marbles become occluded in the duration of the trial, either by the rat or bedding, and reappears, it can thusly be identified using the last known location. Algorithm 10.1 summarizes the marble-tracking procedure. At each frame, the tracking algorithm receives the predicted marble centroids generated by the segmentation model and associates them with existing marble tracks based on spatial proximity. Identity assignment is performed using a two-stage nearest-neighbor matching procedure with progressively relaxed distance thresholds to maintain consistent tracking while minimizing incorrect associations.

During the initial matching stage, each detected marble is assigned the identifier of the nearest tracked marble from the preceding frame, provided that the Euclidean distance be-

tween the two centroids does not exceed 14 pixels. This conservative threshold is intended to preserve identity consistency for marbles exhibiting relatively small inter-frame displacements. Following the first pass, any remaining unmatched detections are processed using a second matching stage in which the allowable association distance is increased to 20 pixels. This relaxed threshold enables the recovery of marble identities in situations where larger apparent displacements occur, such as those resulting from rapid movement, temporary occlusion, or minor localization inaccuracies in the segmentation model.

The resulting tracking procedure produces a temporally consistent, frame-level record of each marble's identity and centroid position. This information forms the basis for subsequent analyses of marble trajectories and rat-marble interactions.

10.4 Heuristic frame-level marble-interaction labeling

Given the predicted rat keypoints and marble locations, a keypoint-based heuristic was developed to identify frames in which the rat was interacting with a marble. This heuristic differs from the NORT interaction heuristic as the objects can be moved and more easily occluded by the rat's body. In short, the heuristic takes the nose position from the current frame and compares it to the marbles detected in an interval of previous frames. This process is done on a frame by frame basis and does not require marble interaction to be sustained for a certain number of frames.

For each frame t, let n_t denote the predicted nose location and $m_{k,t}$ denote the estimated center of marble k. The distance from the rat's nose to marble k is computed as

$$d_{k,t} = \|n_t - m_{k,t}\|_2.$$

The nearest marble in frame t is then identified as

$$k_t^* = \arg\min_k d_{k,t}.$$

A frame is considered a candidate marble-interaction frame if the distance from the nose

Algorithm 10.1 Marble tracking algorithm for maintaining marble identities across frames.

Input: Ordered frames $F_1, \ldots, F_T$; marble detections for each frame; primary matching distance $\tau_{\text{track}} = 14$, secondary matching distance $\upsilon_{\text{track}} = 20$
Output: Frame-level marble IDs and center positions
Initialize marble IDs in first frame, $m_1, \ldots, m_{20}$
Initialize *last_known*, dictionary such that keys are marble ids and values are their locations in the first frame
Initialize *tracker*: $\{F_1 : \{m_1 : (x_{1,1}, y_{1,1}), m_2 : (x_{2,1}, y_{2,1}), ...(x_{20,1}, y_{20,1})\}$
for $t = 2$ to T **do**
    Initialize *found_keys*, dictionary such that keys are marble ids and values are $False$
    Initialize *unfound*, empty list
    Get marble centers in frame F_t: $\hat{m}_{1,t}, \hat{m}_{2,t}, \ldots, \hat{m}_{J,t}$
    for $\hat{m}_{i,t}$ in F_t **do**
        Compute pairwise Euclidean distances between $\hat{m}_{i,t}$ and *last_known*
        $nearest_{id}$, $nearest_{dist} \leftarrow$ last known nearest marble to $\hat{m}_{i,t}$
        if $nearest_{dist} < \tau_{\text{track}}$ **then**
            $tracker[t][nearest_{id}] \leftarrow \hat{m}_{i,t}$
            $found_keys[nearest_{id}] \leftarrow True$
        else
            append $\hat{m}_{i,t}$ to *unfound*
        end if
    end for
    for $\hat{m}_{i,t}$ in *unfound* **do**
        for *key* in *found_keys* where $found_keys[key]$ is $False$ **do**
            if $dist(\hat{m}_{i,t}, last_known[key]) < \upsilon_{track}$ **then**
                $tracker[t][key] \leftarrow \hat{m}_{i,t}$
                $found_keys[key] \leftarrow True$
            end if
        end for
    end for
    update marble locations in *last_known* with updated locations from $tracker[t]$
end for
return *tracker*: $\{F_1 : \{m_1 : (x_{1,1}, y_{1,1}), m_2 : (x_{2,1}, y_{2,1}), ...\}, ...F_T : \{m_1 : (x_{1,T}, y_{1,T}), m_2 : (x_{2,T}, y_{2,T}), ...\}\}$

to the nearest marble is less than a predefined threshold $\tau_{\text{marble}} = 20$:

$$\min_k d_{k,t} \leq \tau_{\text{marble}}.$$

Because a rat may cover a marble with its snout while interacting with a marble, an interval of past frames are considered as well. Specifically, a frame is labeled as `Marble Interaction` if within an interval of past frames, the nose has been within τ_{marble} pixels of a marble.

The values for τ_{marble} and the number of past frames were selected using a grid search. Values tested for τ_{marble} were: $[5, 10, 15, 20, 25, 30]$ pixels. Previous frame intervals tested were: $[5, 10, 15, 20, 30]$. The results of this grid search are outlined in Table 10.4 as precision, recall and F1 score. Based on these results, τ_{marble} was set to 20 and the past frame interval was set to 30. Since the MBT videos are recorded at 15 frames per second, this corresponds to approximately 2 seconds of potential proximity.

Figure 10.6 illustrates the heuristic procedure. Raw frames are passed through the rat keypoint model and the marble segmentation model. The predicted nose, tail-base, and marble centers are then combined to determine whether the rat is near a marble and whether the frame should be labeled as `Marble Interaction`.

Additional heuristics were used to identify `Standing` behavior. `Standing` can occur when the rat rears upward, causing the projected distance between the *nose* and *tail-base* to become compressed in the overhead view. Therefore, frames with a nose–tail-base distance substantially smaller than the average body length are candidate `Standing` frames. The average body length is calculated on a per-video basis, denoted as $\bar{d_V}$ for a video V and this value is then used as comparison for standing. Specifically, a frame t in video V is marked as standing if the following is true:

$$\frac{\|n_t - t_t\|_2}{\bar{d_V}} < c$$

where c is the compression value. As c approaches zero, the closer the nose and tail must be to be considered standing; $c = 0$ would imply that the nose and tail are the same

Past Interval	τ_{marble}	Marble Interaction Pr	Re	F1
5	5	0.249	0.293	0.269
	10	0.262	0.374	0.308
	15	0.267	0.445	0.334
	20	0.269	0.501	0.350
	25	0.260	0.561	0.356
	30	0.243	0.622	0.350
10	5	0.257	0.354	0.298
	10	0.279	0.485	0.354
	15	0.277	0.565	0.372
	20	0.276	0.624	0.383
	25	0.267	0.674	0.382
	30	0.257	0.726	0.379
15	5	0.263	0.389	0.314
	10	**0.288**	0.547	0.377
	15	0.285	0.634	0.393
	20	0.282	0.696	0.401
	25	0.272	0.742	0.398
	30	0.264	0.787	0.395
20	5	0.262	0.400	0.317
	10	**0.288**	0.574	0.383
	15	0.285	0.671	0.400
	20	0.282	0.737	0.408
	25	0.274	0.784	0.406
	30	0.268	0.829	0.405
30	5	0.266	0.425	0.327
	10	**0.288**	0.611	0.391
	15	0.284	0.713	0.407
	20	0.283	0.789	**0.416**
	25	0.276	0.837	**0.416**
	30	0.270	**0.874**	0.413

Table 10.4: *Heuristic rule value selection. The values for the past frame interval and* τ_{marble} *were selected using a grid search. For the find heuristic model, a past interval of 30 frames and* τ_{marble} *of 20 were selected.*

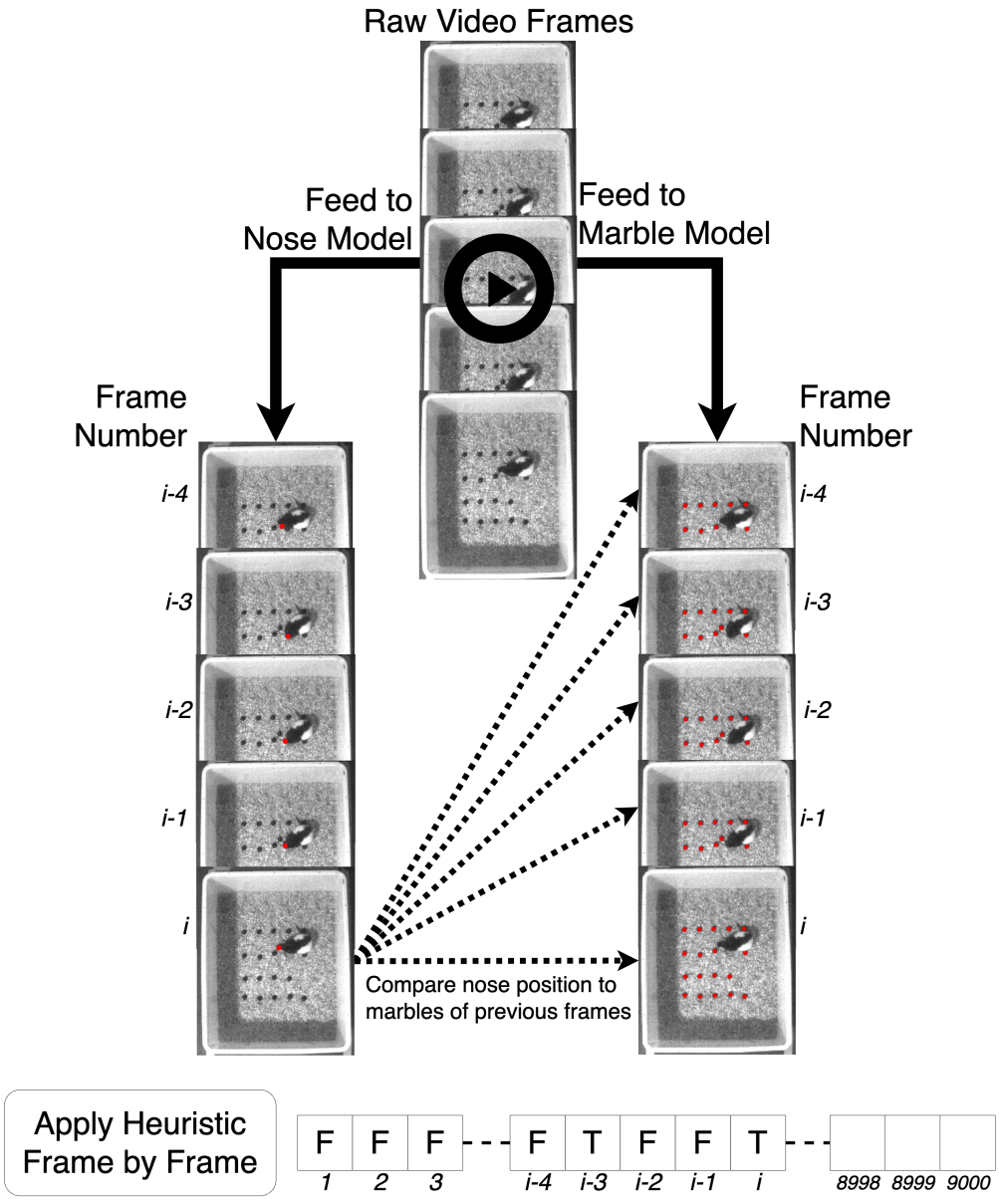


Figure 10.6: *Heuristic procedure for marble-interaction labeling. Raw video frames are processed by the rat keypoint model and marble segmentation model. Predicted nose, tail-base, and marble locations are then used to determine whether the rat is in sustained proximity to the same marble over a short temporal window.*

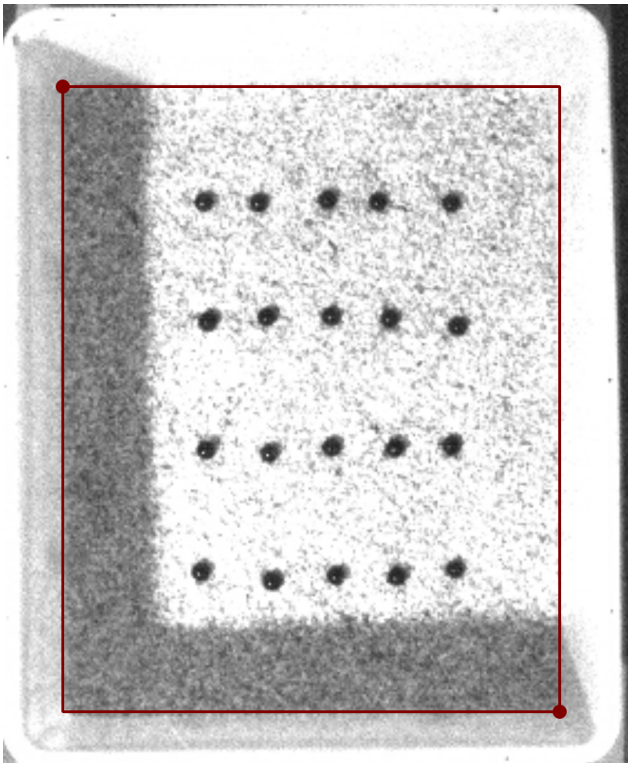


Figure 10.7: *Arena standing boundary. For the instance of standing against the arena boundary, the ideal annotation is where the bedding meets the arena walls, shown in red.*

point. It was determined that $c = 0.1$ was best in terms of trade off between specificity and sensitivity.

`Standing` can also occur when the rat rears against the arena wall. In those cases, the nose keypoint may appear outside or near the boundary of a user-defined arena region. Frames satisfying this wall-rearing condition are also labeled as `Standing`. For standing against the arena wall, an arena boundary must be predefined. The ideal boundary is where the bedding meets the arena walls, which can be seen in Figure 10.7.

When multiple heuristic conditions are satisfied in the same frame, labels are assigned using a consistent precedence rule. In the current analysis, `Marble Interaction` always takes precedence over `Standing`. All remaining frames not labeled `Marble Interaction` or `Standing` are assigned the `Other` label. Further results for the heuristic approach are shown in Table 10.8 to illustrate a comparison between protocols.

10.5 Manual behavior annotation

To evaluate the heuristic and train frame-level behavior classification models, a set of MBT videos was manually annotated using CVAT (CVAT.ai Corporation, 2023) and the same general frame-level annotation framework used for the NORT behavior dataset. Each annotated frame was assigned one of three mutually exclusive labels: `Marble Interaction`, `Standing`, or `Other`.

The `Marble Interaction` label was used for frames in which the rat appeared to actively investigate, contact, or otherwise direct behavior toward a marble. The `Standing` label was used for frames in which the rat displayed an upright or rearing posture, including wall-directed standing. The `Other` label was used for all remaining frames, including ordinary locomotion, non-marble-directed exploration, grooming, and any low-mobility states not explicitly separated into resting or freezing categories.

In isolation, many frames can be ambiguous due to the resolution as well as the nature of the rat. For example, rats may pick up a marble and sit in a hunched position while examining the marble. In cases like these, annotators must refer to the surrounding frames to infer further context. As a general rule, preference was given to marble interaction when other behaviors occurred simultaneously. Using this protocol, the entirety of each video within the dataset was annotated in a temporally consistent and context-aware manner.

Table 10.5 summarizes the current MBT behavior annotation dataset. The dataset is imbalanced, with `Other` being the most frequent label. This imbalance reflects the fact that marble-directed behavior and standing occur during relatively short portions of the trial, whereas most frames contain other movement or non-target behavior. For this reason, evaluation of the MBT behavior models should include class-level precision, recall, and F1-score in addition to overall accuracy. Additionally, `Other` was undersampled for the training dataset with only 1,550 frames taken from four training video for a total of 6,200 frames.

Subset	Video	Marb. Int.	Standing	Other	Total
Train	47.07	3,909	1,713	1,500	7,122
	48.06	1,210	2,330	1,500	5,040
	49.06	590	673	1,500	2,763
	50.07	944	1,463	1,500	3,907
	64.02	389	933	0	1,322
	75.01	92	163	0	255
	Total	7,134	7,275	6,200	20,609
Val	47.06	2,539	1,143	5,189	8,871
	50.05	271	987	6,348	7,606
	Total	2,810	2,130	11,537	16,477
Test	77.05	1,310	1,305	6,590	9,205
	81.08	859	1,038	7,211	9,108
	84.06	2,545	1,796	5,078	9,419
	Total	4,714	4,139	18,879	27,732
All	**Total**	14,658	13,544	36,616	64,818

Table 10.5: *MBT behavior annotation dataset used for frame-level marble-interaction classification. Labels are mutually exclusive and include Marble Interaction, Standing, and Other. Due to the over-abundance of the Other class, it was undersampled in training.*

10.6 YOLOv11 behavior classification for MBT

In addition to the keypoint-based heuristic, YOLOv11 classification models were trained to predict frame-level MBT behavior labels directly from image frames. This follows the same general strategy used for NORT behavior classification, where frame-level classifiers were trained to distinguish task-relevant behavior classes from other activity. For the MBT videos, the classification task was defined using the three labels summarized in Table 10.5: `Marble Interaction`, `Standing`, and `Other`.

The initial model-selection process used a broad screening step to compare multiple YOLOv11 classification model sizes. The goal of this step was to identify an appropriate trade-off between validation performance and computational cost. Because the MBT dataset is imbalanced and `Marble Interaction` is the primary behavior of interest, model selection should not rely on accuracy alone. Instead, validation macro-F1 and class-level F1-scores, especially `Marble Interaction` precision and recall, should be used to evaluate candidate models.

Based on the initial screening, the nano and xlarge YOLOv11 configurations were selected

Model	Accuracy	Macro-F1	Marb. Int. F1	Standing F1	Other F1
YOLOv11-nano	**0.744**	**0.619**	0.455	**0.566**	**0.837**
YOLOv11-small	0.686	0.529	0.336	0.454	0.796
YOLOv11-medium	0.491	0.438	0.278	0.431	0.606
YOLOv11-large	0.694	0.542	0.330	0.495	0.802
YOLOv11-xlarge	0.678	0.557	**0.474**	0.408	0.789

Table 10.6: *Validation performance of YOLOv11 classification models of varying sizes for MBT behavior classification. Models nano and xlarge were selected for further optimization based on validation macro-F1 and Marble Interaction performance.*

for further optimization. These models were then trained more extensively while varying optimization settings such as learning rate, optimizer, batch size, number of epochs, and initialization strategy. The selected configuration was chosen based on validation macro-F1 while also considering recall for Marble Interaction. Table 10.7 outlines the results of varying the optimization settings, reporting precision, recall, and F1-score for each class as well as a macro-average.

Given the validation results, the YOLOv11-`nano` model trained with a batch size of 32, the Adamax optimizer, and a learning rate of 0.0003 was selected as the final MBT behavior classification model. This model was further evaluated and compared against the heuristic model by evaluating on three held out videos. A hybrid approach was also implemented. In this hybrid, when the class labels from the YOLO model and heuristic model match, apply the label. In the case that the labels do not match, then apply the label of `Other` to the frame. This methodology contrasts to the NORT-hybrid protocol as both the heuristic and YOLOv11 models over-classified `Standing` and `Marble Interaction`

The class-level precision, recall, and F1-scores for each approach, together with their respective macro-averages, are reported in Table 10.8.

The YOLOv11 classifier and the keypoint-based heuristic provide complementary approaches to MBT behavior annotation. The heuristic is transparent and directly tied to interpretable spatial conditions involving the rat and the marbles. However, it depends on reliable keypoint and marble localization and may struggle with visually ambiguous behaviors. The YOLOv11 classifier can learn additional visual cues directly from the image frame, but it requires sufficient labeled examples and may be less interpretable. Comparing these approaches provides insight into whether MBT marble-interaction analysis is better

Model	**Other**			**Marble Interaction**			**Standing**			**Macro-avg**		
size-batch-optimizer-lr	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
n-16-Adam-1e-05	0.683	0.816	0.743	0.436	0.400	0.418	0.849	0.481	0.614	0.656	0.566	0.608
n-16-Adam-0.0001	0.809	0.838	0.823	0.436	0.512	0.471	0.866	0.628	0.728	0.704	0.659	0.681
n-16-Adam-0.0003	**0.902**	0.802	0.849	0.183	0.430	0.256	0.869	0.800	0.833	0.651	0.677	0.664
n-16-Adam-0.0005	0.795	0.838	0.816	0.430	0.424	0.427	0.895	0.713	0.794	0.707	0.658	0.682
n-16-Adam-0.001	0.779	0.822	0.800	0.375	0.379	0.377	0.881	0.682	0.769	0.678	0.627	0.652
n-16-Adamax-1e-05	0.665	0.828	0.737	0.490	0.379	0.427	0.867	0.516	0.647	0.674	0.574	0.620
n-16-Adamax-0.0001	0.854	0.795	0.824	0.190	0.366	0.250	0.851	0.693	0.764	0.632	0.618	0.625
n-16-Adamax-0.0003	0.812	0.832	0.822	0.407	0.450	0.427	0.880	0.701	0.780	0.700	0.661	0.680
n-16-Adamax-0.0005	0.858	0.830	0.844	0.355	0.501	0.415	0.878	0.732	0.799	0.697	0.688	0.692
n-16-Adamax-0.001	0.842	0.837	0.839	0.396	0.556	0.462	0.885	0.658	0.755	0.708	0.684	0.695
n-32-Adam-1e-05	0.696	0.830	0.757	0.428	0.362	0.392	0.886	0.542	0.672	0.670	0.578	0.621
n-32-Adam-0.0001	0.737	0.857	0.793	0.548	0.421	0.476	0.851	0.625	0.720	0.712	0.634	0.671
n-32-Adam-0.0003	0.846	0.816	0.831	0.271	0.463	0.342	0.885	0.656	0.753	0.667	0.645	0.656
n-32-Adam-0.0005	0.767	0.837	0.800	0.470	0.344	0.397	0.846	**0.870**	**0.858**	0.694	0.684	0.689
n-32-Adam-0.001	0.785	0.823	0.803	0.386	0.418	0.401	0.891	0.657	0.756	0.687	0.633	0.659
n-32-Adamax-1e-05	0.575	0.812	0.674	0.477	0.370	0.416	0.894	0.407	0.559	0.649	0.530	0.583
n-32-Adamax-0.0001	0.822	0.832	0.827	0.387	0.464	0.422	0.878	0.684	0.769	0.696	0.660	0.677
n-32-Adamax-0.0003	0.831	0.858	0.844	0.500	0.514	**0.507**	0.859	0.715	0.780	0.730	0.695	**0.712**
n-32-Adamax-0.0005	0.856	0.839	0.848	0.389	0.506	0.440	0.884	0.741	0.806	0.710	0.695	0.702
n-32-Adamax-0.001	0.860	0.803	0.831	0.197	0.385	0.261	0.904	0.718	0.800	0.654	0.636	0.644
n-64-Adam-1e-05	0.836	0.816	0.826	0.288	0.389	0.331	0.835	0.689	0.755	0.653	0.631	0.642
n-64-Adam-0.0001	0.813	0.826	0.819	0.381	0.411	0.395	0.853	0.721	0.781	0.682	0.653	0.667
n-64-Adam-0.0003	0.872	0.833	**0.852**	0.336	**0.599**	0.431	0.903	0.680	0.775	0.704	**0.704**	0.704
n-64-Adam-0.0005	0.798	0.828	0.813	0.383	0.387	0.385	0.890	0.734	0.804	0.690	0.650	0.669
n-64-Adam-0.001	0.752	0.838	0.793	0.522	0.385	0.444	0.800	0.738	0.768	0.691	0.654	0.672
n-64-Adamax-1e-05	0.630	0.819	0.712	0.483	0.391	0.432	0.886	0.457	0.603	0.666	0.556	0.606
n-64-Adamax-0.0001	0.767	0.821	0.794	0.412	0.418	0.415	0.854	0.621	0.720	0.678	0.620	0.648
n-64-Adamax-0.0003	0.831	0.817	0.824	0.303	0.504	0.378	0.888	0.618	0.729	0.674	0.647	0.660
n-64-Adamax-0.0005	0.843	0.815	0.829	0.288	0.459	0.354	0.887	0.680	0.770	0.673	0.651	0.662
n-64-Adamax-0.001	0.789	0.844	0.815	0.452	0.483	0.467	0.901	0.628	0.740	0.714	0.652	0.681
x-16-Adam-1e-05	0.833	0.810	0.821	0.265	0.425	0.327	0.884	0.658	0.755	0.661	0.631	0.646
x-16-Adam-0.0001	0.810	0.853	0.831	0.494	0.501	0.497	0.899	0.695	0.784	0.734	0.683	0.708
x-16-Adam-0.0003	0.752	0.840	0.793	0.491	0.383	0.431	0.881	0.739	0.804	0.708	0.654	0.680
x-16-Adam-0.0005	0.766	0.858	0.810	0.553	0.454	0.499	0.860	0.665	0.750	0.727	0.659	0.691
x-16-Adam-0.001	0.777	0.826	0.801	0.395	0.393	0.394	0.894	0.679	0.772	0.689	0.633	0.660
x-16-Adamax-1e-05	0.813	0.827	0.820	0.376	0.498	0.429	0.885	0.628	0.735	0.692	0.651	0.671
x-16-Adamax-0.0001	0.869	0.812	0.840	0.284	0.475	0.356	0.855	0.740	0.793	0.669	0.676	0.673
x-16-Adamax-0.0003	0.828	0.837	0.833	0.414	0.464	0.437	0.894	0.744	0.812	0.712	0.682	0.696
x-16-Adamax-0.0005	0.817	0.843	0.830	0.461	0.467	0.464	0.870	0.735	0.797	0.716	0.682	0.698
x-16-Adamax-0.001	0.860	0.797	0.827	0.203	0.320	0.248	0.854	0.810	0.831	0.639	0.642	0.640
x-32-Adam-1e-05	0.852	0.812	0.832	0.233	0.438	0.304	0.894	0.662	0.761	0.660	0.637	0.648
x-32-Adam-0.0001	0.815	0.836	0.825	0.422	0.500	0.458	0.871	0.650	0.745	0.703	0.662	0.682
x-32-Adam-0.0003	0.819	0.836	0.827	0.400	0.553	0.464	0.910	0.616	0.735	0.710	0.668	0.688
x-32-Adam-0.0005	0.859	0.797	0.827	0.194	0.345	0.249	0.871	0.753	0.807	0.641	0.632	0.637
x-32-Adam-0.001	0.739	0.814	0.775	0.378	0.397	0.387	0.894	0.572	0.698	0.670	0.594	0.630
x-32-Adamax-1e-05	0.730	0.838	0.780	0.476	0.392	0.430	0.909	0.644	0.754	0.705	0.625	0.662
x-32-Adamax-0.0001	0.819	0.846	0.832	0.459	0.434	0.446	0.883	0.807	0.843	0.721	0.696	0.708
x-32-Adamax-0.0003	0.815	0.835	0.825	0.400	0.526	0.454	**0.912**	0.630	0.745	0.709	0.664	0.686
x-32-Adamax-0.0005	0.837	0.832	0.835	0.373	0.448	0.407	0.897	0.755	0.820	0.702	0.678	0.690
x-32-Adamax-0.001	0.872	0.811	0.841	0.230	0.479	0.311	0.908	0.710	0.796	0.670	0.666	0.668
x-64-Adam-1e-05	0.812	0.818	0.815	0.266	0.382	0.313	0.900	0.625	0.738	0.659	0.608	0.633
x-64-Adam-0.0001	0.855	0.832	0.844	0.370	0.531	0.436	0.871	0.697	0.775	0.699	0.687	0.693
x-64-Adam-0.0003	0.775	**0.863**	0.816	**0.572**	0.453	0.506	0.870	0.721	0.789	**0.739**	0.679	0.708
x-64-Adam-0.0005	0.836	0.818	0.827	0.322	0.381	0.349	0.873	0.804	0.837	0.677	0.668	0.672
x-64-Adam-0.001	0.755	0.817	0.785	0.377	0.324	0.348	0.884	0.737	0.804	0.672	0.626	0.648
x-64-Adamax-1e-05	0.745	0.848	0.793	0.522	0.451	0.484	0.899	0.619	0.733	0.722	0.640	0.678
x-64-Adamax-0.0001	0.864	0.816	0.840	0.272	0.540	0.361	0.894	0.669	0.765	0.677	0.675	0.676
x-64-Adamax-0.0003	0.864	0.827	0.845	0.295	0.449	0.356	0.910	0.753	0.824	0.690	0.676	0.683
x-64-Adamax-0.0005	0.866	0.810	0.837	0.223	0.454	0.300	0.896	0.691	0.780	0.662	0.652	0.657
x-64-Adamax-0.001	0.814	0.796	0.805	0.197	0.289	0.234	0.904	0.694	0.785	0.638	0.593	0.615

Table 10.7: *Validation performance of the selected YOLOv11 MBT behavior classification model. Each model is named in the following convention: (model size) − (batch size) − (optimizer) − (learning rate). Based on these results, the YOLOv11-*`nano` *model, trained with a batch size of 32 and the Adamax optimizer using a learning rate of 0.0003, was selected as the final model.*

Video	Model	Standing			Marble Int.			Other			Macro-avg		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
77.05	Heuristic Model	0.594	0.253	0.355	0.268	0.546	0.360	0.793	0.720	0.755	0.552	0.506	0.490
	n-32-Adamax-0.0003	0.181	0.841	0.298	0.316	0.703	0.436	0.819	0.028	0.054	0.439	0.524	0.263
	Hybrid Model	0.612	0.245	0.350	0.370	0.373	0.371	0.768	0.858	0.811	0.583	0.492	0.511
81.08	Heuristic Model	0.755	0.782	0.768	0.364	**0.824**	0.505	0.939	0.792	0.859	0.686	**0.800**	0.711
	n-32-Adamax-0.0003	0.210	0.914	0.342	0.335	0.815	0.475	0.918	0.319	0.473	0.488	0.683	0.430
	Hybrid Model	**0.793**	0.760	**0.776**	0.521	0.665	**0.584**	0.923	**0.898**	**0.911**	**0.746**	0.774	**0.757**
84.06	Heuristic Model	0.634	0.921	0.751	0.458	0.693	0.551	0.765	0.446	0.563	0.619	0.686	0.622
	n-32-Adamax-0.0003	0.243	0.994	0.391	0.647	0.511	0.571	**0.947**	0.014	0.028	0.613	0.506	0.330
	Hybrid Model	0.634	**0.917**	0.749	**0.701**	0.464	0.558	0.744	0.753	0.748	0.693	0.711	0.685
Macro-avg	Heuristic Model	0.661	0.652	0.625	0.363	0.688	0.472	0.832	0.653	0.726	0.619	0.664	0.607
	n-32-Adamax-0.0003	0.212	0.917	0.344	0.433	0.676	0.494	0.895	0.120	0.185	0.513	0.571	0.341
	Hybrid Model	0.680	0.641	0.625	0.530	0.501	0.505	0.812	0.836	0.823	0.674	0.659	0.651

Table 10.8: *Class-level test performance for MBT behavior classification.*

supported by explicit spatial reasoning, direct frame classification, or a hybrid strategy that combines both.

10.7 Preliminary discussion and limitations

The MBT analysis demonstrates that the automated behavior annotation framework developed in this dissertation can be extended to a more challenging object-interaction setting. Compared with NORT, the MBT setting introduces several additional sources of difficulty: lower-resolution videos, small objects, repeated object instances with similar appearance, object displacement over time, and partial or complete occlusion caused by bedding. These challenges make marble-interaction analysis more difficult than fixed-object interaction analysis and motivate the combination of keypoint detection, marble segmentation, identity tracking, and temporal heuristics.

The current MBT extension should be interpreted as a preliminary but important step toward richer automated analysis of marble-burying behavior. Rather than relying only on the number of marbles buried at the end of the trial, the proposed workflow provides a way to recover temporal information about when and where marble-directed interactions occur. In addition, persistent marble tracking makes it possible to examine marble displacement patterns and to relate those displacements to rat behavior over the course of the trial.

Several limitations remain. First, the marble segmentation model was trained using a relatively small number of annotated frames from a limited set of videos. Additional

annotated examples from more videos would likely improve robustness, especially for marbles that are partially buried or difficult to distinguish from the bedding. Second, tracking individual marbles is challenging when marbles are occluded, displaced rapidly, or visually indistinguishable from one another. Finally, the current frame-level behavior labels do not directly distinguish between different forms of marble-directed behavior, such as sniffing, digging, pushing, and burying.

Future work should expand the MBT annotation dataset, validate marble trajectories against manual review, and compare automated frame-level interaction measures with traditional endpoint marble-burying counts. Additional temporal models could also be explored to capture longer behavioral sequences and distinguish sustained burying bouts from brief incidental contact. Rat orientation can also be incorporated using the vector from the tail-base to the nose as an estimate of the rat's body direction. A frame can then be required to satisfy a certain orientation threshold, indicating that the rat is oriented toward the marble rather than merely passing beside it, to annotate a frame as marble interaction. Despite these remaining challenges, the MBT extension illustrates the adaptability of the proposed framework and provides a foundation for automated, temporally resolved analysis of marble-directed behavior in rodent trials.

Chapter 11

Conclusions and future work

11.1 Conclusions

This dissertation presented a scalable framework for automated behavioral annotation in rodent trials, with a primary focus on the Novel Object Recognition Test (NORT) and extensions to additional behavioral paradigms, including Fisher 344 NORT, Elevated Plus Maze (EPM), and marble-interaction analysis in the Marble Burying Test (MBT). The work addressed a central challenge in behavioral neuroscience: the need for efficient, objective, and reproducible methods for analyzing large volumes of behavioral video data. Manual annotation remains valuable because it incorporates domain expertise, but it is time-intensive, difficult to scale, and susceptible to observer variability, especially when behavioral events are brief, ambiguous, or occur near transition boundaries. As behavioral neuroscience experiments continue to generate increasingly large video datasets, automated systems are becoming essential for supporting reproducible and large-scale quantitative analysis.

The central computational contribution of this dissertation is a hybrid, modular approach in which learned visual perception is combined with explicit spatial representations of the experimental task. Rather than requiring a single model to implicitly learn both the visual appearance of a behavior and the geometry needed to interpret that behavior, the framework assigns these roles to complementary components. YOLOv11-based models provide keypoint

localization and frame-level visual recognition, while spatially informed heuristics incorporate animal orientation, posture, object proximity, and arena geometry. This decomposition allows learned representations to be shared where appropriate while preserving interpretable, task-specific behavioral definitions.

The core methodological contribution was demonstrated in the NORT experiments, where generic behavioral recognition was separated from object-specific spatial interpretation. A unified classifier trained across the 2-object and 5-object configurations identifies frame-level behavioral states such as Object Interaction, Standing, and Other, while nose and tail-base keypoints provide the spatial information needed to determine the specific object involved. This design avoids requiring a separate behavioral class for each combination of behavior and object configuration and permits training data from related NORT settings to contribute to a shared visual classifier.

The NORT experiments demonstrated the value of combining information across experimental configurations. Models trained jointly on both 2-object and 5-object videos achieved stronger overall generalization than models trained separately on individual configurations. This finding suggests that many visual cues associated with object exploration and standing are shared across NORT settings, even when the number and arrangement of objects differ. The combined model also reduced the need to train separate behavior classifiers for each arena layout, improving the scalability of the approach. To retain object-specific interpretability, generic object-interaction predictions from the combined classifier were paired with keypoint-based spatial reasoning to determine which specific object the rat was interacting with. This strategy enabled the framework to benefit from pooled training data while still producing analysis-ready outputs needed by behavioral neuroscientists.

The results further demonstrate the utility of a compact pose representation as the interface between learned perception and task-specific spatial reasoning. Accurate detection of the nose and tail-base supported downstream behavior classification, object-interaction localization, trajectory analysis, heatmap generation, and ethogram construction. Interpolation and post-processing further improved the usability of predicted keypoint trajectories by reducing the effect of brief missing detections. These results indicate that a relatively com-

pact keypoint representation can support multiple behavioral analysis tasks when combined with task-specific spatial reasoning.

Beyond model development, this dissertation demonstrated the practical value of automated annotation for large-scale behavioral analysis. The final NORT workflow was applied to approximately 500 videos, enabling analysis of object exploration, novel object recognition, place recognition, rearing behavior, and immobility-related behaviors across experimental groups and timepoints. Such analysis would be prohibitively time-consuming using fully manual annotation. By substantially reducing the amount of human labor required for behavioral coding, the proposed system allows researchers to process larger datasets more consistently while preserving expert involvement in behavioral interpretation and validation.

The dissertation also showed that the proposed framework can be adapted beyond the original NORT setting. For Fisher 344 NORT videos, the keypoint and behavior classification components were adapted to a new animal appearance, demonstrating the importance of fine-tuning when visual characteristics differ from the original training data. For EPM analysis, the framework was adapted from object-centered interaction detection to region-based spatial analysis, where keypoints and arena definitions were used to estimate time spent in open arms, closed arms, and the center region, as well as arm-entry behavior. For marble-interaction analysis, the framework was extended to a more challenging dynamic-object setting in which multiple small marbles must be detected, tracked, and related to rat keypoints over time. These extensions illustrate that the general workflow is not limited to one behavioral paradigm, but can be modified to support different experimental goals, spatial layouts, and behavioral definitions.

Viewed relative to prior automated behavioral-analysis approaches, the contribution of this dissertation is therefore not pose estimation, direct visual behavior classification, or rule-based behavioral inference individually. Pose-estimation frameworks provide detailed localization, systems such as DeepEthogram learn behavioral labels from visual data, BehaviorDEPOT demonstrates the interpretability of pose-derived behavioral rules, and EXPLORE demonstrates learned classification of object exploration. The framework developed in this dissertation integrates these complementary strategies and, importantly, separates

generic learned behavioral recognition from experiment-specific spatial interpretation. The NORT experiments show that this decomposition can support sharing across related experimental configurations, while the Fisher 344, EPM, and MBT extensions demonstrate both the opportunities and limitations of transferring the resulting components across changes in animal appearance, arena geometry, objects, and behavioral objectives. The broader contribution is therefore a modular strategy for deciding what should be learned from data, what should be represented explicitly from experimental structure, and what must be adapted when the experimental domain changes.

11.2 Future work

Despite the strong results achieved in this dissertation, several limitations remain that should be addressed in future iterations of this work. First, the behavior classifiers operate primarily at the frame level. Although frame-level classification is useful and computationally efficient, it does not explicitly model longer temporal dependencies. Many behavioral ambiguities occur near transition boundaries, where even human annotators may disagree about the exact onset or offset of a behavior. Incorporating temporal models, such as temporal convolutional networks, recurrent neural networks, transformer-based video models, or post-hoc sequence smoothing, could improve consistency and reduce brief prediction fluctuations.

Second, although the NORT dataset includes both 2-object and 5-object configurations, the amount of manually annotated behavior data remains limited relative to the full diversity of possible experimental conditions. Generalization may be affected by changes in lighting, camera placement, arena geometry, object appearance, animal strain, coat color, and posture distribution. Additional annotated data from more laboratories, rat strains, object sets, and recording conditions would strengthen the robustness of the framework and support broader deployment.

An important related direction is to quantify the annotation cost required to adapt the framework to a new domain. The Fisher 344 experiments suggest that different pipeline components may require different amounts and types of adaptation: keypoint localization

can recover strongly after targeted fine-tuning, whereas behavior classification must also account for differences in behavioral frequency, posture, and expression. Future studies should therefore construct annotation-efficiency curves that measure performance as a function of the number and diversity of newly labeled frames. Such experiments could provide practical guidance regarding when an existing model can be reused, when targeted fine-tuning is sufficient, and when substantial re-annotation or model redesign is required for a new strain, laboratory, camera configuration, or behavioral paradigm.

A complementary direction is to reduce the amount of manual configuration required by the spatial-reasoning component itself. The current framework benefits from explicit knowledge of object locations, arena boundaries, and assay-specific regions, but these definitions must be adapted when the experimental setup changes. Future work could automatically detect or register arena geometry and experimental objects from video, allowing the task-specific spatial layer to be initialized with minimal manual intervention. Combining automatic scene understanding with the existing interpretable reasoning framework could further improve portability across laboratories and experimental configurations.

Third, the current behavior label space is intentionally focused on a limited set of behaviors needed for the present studies. For NORT, the primary labels include `Object Interaction`, `Standing`, and `Other`, with additional post-processing used to generate object-specific interaction measures and analysis-ready summaries. However, rodent behavior is richer and more continuous than any small set of discrete labels can fully capture. Future work could extend the framework to include additional behaviors, continuous behavioral state estimation, movement dynamics, bout-level analysis, and higher-level behavioral sequences.

Fourth, the EPM and MBT components should be interpreted as extensions of the core NORT framework rather than as fully mature standalone systems. The EPM extension demonstrates how keypoints and arena-region definitions can support region occupancy and arm-entry analysis, but additional validation would be needed for more subtle EPM behaviors such as head dips, risk assessment, and stretch-attend postures. Similarly, the MBT extension introduces marble detection, marble tracking, and frame-level marble-interaction analysis, but further work is needed to validate marble trajectories, distinguish different

types of marble-directed behavior, and compare automated temporal measures with traditional endpoint marble-burying scores.

Future work will also explore more interactive annotation workflows. Rather than replacing human annotators entirely, automated systems can be used to pre-label videos, identify uncertain segments, and direct expert attention to frames or bouts that require review. Active learning and human-in-the-loop correction could reduce annotation effort while improving model performance over time. This would be especially valuable for behaviors that are rare, ambiguous, or highly dependent on expert judgment.

Finally, reproducibility and accessibility remain important goals. The planned release of code, trained models, annotation tools, and a subset of the annotated NORT data will help address the shortage of publicly available rat behavior datasets and benchmarks. These resources can support future comparison studies, model improvements, and adaptation of the framework to additional behavioral assays.

Overall, this dissertation demonstrates that scalable rodent behavior annotation benefits from combining learned visual perception with explicit, pose-informed representations of experimental structure. The proposed system achieves scalable and interpretable analysis of NORT videos and provides a foundation for extension to other behavioral paradigms. Beyond its direct application to NORT and fragile X syndrome-related studies, this work contributes to the broader development of reliable, reproducible, and scalable computational tools for behavioral phenotyping in preclinical neuroscience.

Bibliography

D. J. Anderson and P. Perona. Toward a science of computational ethology. *Neuron*, 84(1): 18–31, 2014.

M. Becegato and R. H. Silva. Object recognition tasks in rats: Does sex matter? *Frontiers in behavioral neuroscience*, 16:970452, 2022. doi: 10.3389/fnbeh.2022.970452. URL https://doi.org/10.3389/fnbeh.2022.970452.

J. P. Bohnslav, N. K. Wimalasena, K. J. Clausing, Y. Y. Dai, D. A. Yarmolinsky, T. Cruz, A. D. Kashlan, M. E. Chiappe, L. L. Orefice, C. J. Woolf, and C. D. Harvey. Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife*, 10:e63377, 2021.

A. Borreca, M. De Luca, A. Ferrante, Z. Boussadia, A. Pignataro, A. Martire, and M. Ammassari-Teule. Fmr1-ko mice failure to detect object novelty associates with a post-test decrease of structural and synaptic plasticity upstream of the hippocampus. *Scientific Reports*, 13(1):755, 2023. doi: 10.1038/s41598-023-27991-9. URL https://doi.org/10.1038/s41598-023-27991-9.

CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), Nov. 2023. URL https://github.com/cvat-ai/cvat.

G. De Brouwer, A. Fick, B. H. Harvey, and D. W. Wolmarans. A critical inquiry into marble-burying as a preclinical screening paradigm of relevance for anxiety and obsessive–compulsive disorder: Mapping the way forward. *Cognitive, Affective, & Behavioral Neuroscience*, 19(1):1–39, 2019.

R. M. J. Deacon. Digging and marble burying in mice: simple methods for in vivo identification of biological impacts. *Nat Protoc*, 1(1):122–124, 2006.

X. Du, X. Zhang, T. Li, X. Chen, X. Yu, and H. Wang. YOLO-WAS: A Lightweight Apple Target Detection Method Based on Improved YOLO11. *Agriculture*, 15(14): 1521, July 2025. doi: 10.3390/agriculture15141521. URL http://dx.doi.org/10.3390/agriculture15141521.

S. D'hers, A. D. Robles, S. Ojea Ramos, G. Bollini, and M. Feld. Rainstorm: Automated behavioral analysis for mice exploratory behavior using artificial neural networks. *bioRxiv*, pages 2025–04, 2025.

E. Fazzari, D. Romano, F. Falchi, and C. Stefanini. Animal behavior analysis methods using deep learning: A survey. *arXiv preprint arXiv:2405.14002*, 2024.

C. J. Gabriel, Z. Zeidler, B. Jin, C. Guo, C. M. Goodpaster, A. Q. Kashay, A. Wu, M. Delaney, J. Cheung, L. E. DiFazio, et al. Behaviordepot is a simple, flexible tool for automated behavioral detection based on markerless pose tracking. *Elife*, 11:e74314, 2022.

Y. Gan, X. Ren, H. Liu, Y. Chen, and P. Lin. A novel lightweight yolo11-based framework for precisely locating diverse ship targets in complex optical remote sensing photographs. *Measurement Science and Technology*, 36(4):045409, Apr. 2025. doi: 10.1088/1361-6501/adc75b. URL http://dx.doi.org/10.1088/1361-6501/adc75b.

N. L. Goodwin, J. J. Choong, S. Hwang, K. Pitts, L. Bloom, A. Islam, Y. Y. Zhang, E. R. Szelenyi, X. Tong, E. L. Newman, et al. Simple behavioral analysis (simba) as a platform for explainable machine learning in behavioral neuroscience. *Nature Neuroscience*, pages 1–14, 2024.

J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, 8:e47994, 2019.

A. Hatamian, L. Levine, H. E. Oskouie, and M. Sarrafzadeh. Exploring the impact of dataset statistical effect size on model performance and data sample size sufficiency. 2025. doi: 10.48550/ARXIV.2501.02673. URL https://arxiv.org/abs/2501.02673.

V. Ibañez, L. Bohlen, F. Manuella, I. Mansuy, F. Helmchen, and A.-S. Wahl. Explore: a novel deep learning-based analysis method for exploration behaviour in object recognition tests. *Scientific Reports*, 13(1):4249, 2023.

M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1): 64–67, 2013.

R. Khanam and M. Hussain. Yolov11: An overview of the key architectural enhancements, 2024. URL https://arxiv.org/abs/2410.17725.

L. M. Lueptow. Novel object recognition test for the investigation of learning and memory in mice. *Journal of visualized experiments: JoVE*, (126):55718, 2017.

X. Luo and T. Matsumoto. Tensile strength estimation of uhpfrc based on predicted cracking location using deep learning. *Materials*, 18(10):2237, May 2025. doi: 10.3390/ma18102237. URL http://dx.doi.org/10.3390/ma18102237.

A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

G. Ntoulas, C. Brakatselos, G. Nakas, M.-Z. Asprogerakas, F. Delis, L. J. Leontiadis, G. Trompoukis, C. Papatheodoropoulos, D. Gkikas, D. Valakos, et al. Multi-level profiling of the fmr1 ko rat unveils altered behavioral traits along with aberrant glutamatergic function. *Translational Psychiatry*, 14(1):104, 2024. doi: 10.1038/s41398-024-02815-0. URL https://doi.org/10.1038/s41398-024-02815-0.

T. D. Pereira, N. Tabris, A. Matsliah, D. M. Turner, J. Li, S. Ravindranath, E. S. Papadoyannis, E. Normand, D. S. Deutsch, Z. Y. Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

M. Perez and C. Toler-Franklin. Cnn-based action recognition and pose estimation for classifying animal behavior from videos: A survey. *arXiv preprint arXiv:2301.06187*, 2023.

P. A. Philipsberg, Z. C. Wick, K. S. Diego, N. Vaughan, A. Galas, A. Jurkowski, Y. Feng, L. M. Vetere, L. Chen, I. Soler, et al. Chronotate: An open-source tool for manual timestamping and quantification of animal behavior. *Neuroscience letters*, 814:137461, 2023.

W. T. Powers. A feedback model for behavior: Application to a rat experiment. *Behavioral Science*, 16(6):558–563, 1971.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008.

K. Rutten, O. Reneerkens, H. Hamers, A. Şık, I. McGregor, J. Prickaerts, and A. Blokland. Automated scoring of novel object recognition in rats. *Journal of neuroscience methods*, 171(1):72–77, 2008.

H. D. Schlinger. Behavior analysis and behavioral neuroscience, 2015.

O. Sturman, L. von Ziegler, C. Schläppi, F. Akyol, M. Privitera, D. Slominski, C. Grimm, L. Thieren, V. Zerbi, B. Grewe, et al. Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology*, 45(11):1942–1952, 2020.

A. Thomas, A. Burant, N. Bui, D. Graham, L. A. Yuva-Paylor, and R. Paylor. Marble burying reflects a repetitive and perseverative behavior more than novelty-induced anxiety. *Psychopharmacology*, 204(2):361–373, 2009.

Ultralytics YOLO. Ultralytics yolo: Configuration (train settings). https://docs.ultralytics.com/usage/cfg/, 2025. Accessed Oct. 14, 2025.

E. A. van Dam, L. P. Noldus, and M. A. van Gerven. Deep learning improves automated rodent behavior recognition within a specific experimental setup. *Journal of neuroscience methods*, 332:108536, 2020.

E. A. Van Dam, L. P. Noldus, and M. A. Van Gerven. Disentangling rodent behaviors to improve automated behavior recognition. *Frontiers in Neuroscience*, 17:1198209, 2023.

L. von Ziegler, O. Sturman, and J. Bohacek. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology*, 46(1):33–44, 2021.

L. Wahl, A. M. Punt, T. Arbab, I. Willuhn, Y. Elgersma, and A. Badura. A novel automated approach for improving standardization of the marble burying test enables quantification of burying bouts and activity characteristics. *Eneuro*, 9(2), 2022.

A. A. Walf and C. A. Frye. The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. *Nat Protoc*, 2(2):322–328, 2007.

M. R. Whiteway, E. S. Schaffer, A. Wu, E. K. Buchanan, O. F. Onder, N. Mishra, and L. Paninski. Semi-supervised sequence modeling for improved behavioral segmentation. *bioRxiv*, pages 2021–06, 2021.

Appendix A

Global NORT video analysis

We used the model to annotate all 500+ videos in our collection. The output from the model was used to explore group differences in behavior during the novel object recognition task in rats. Data from the 5-object NORT videos are reported here. A Bayesian model was used to analyze the pattern of object exploration, rearing, and resting/freezing behaviors across the three days of testing, where animals were exposed to objects. Data from the habituation day was not included in the analysis. The model included the genotype-sex group of the animals, day of testing, and the interaction of these two variables as predictors, and the intercept and slope of day as random effects. The model converged well without any divergent transitions, sufficiently large effective sample sizes, and Rhat values of 1.01 or less. Furthermore, visual check of the Markov Chain Monte Carlo (MCMC) sampling also showed good convergence, with overlapping chains in the caterpillar plots. This model revealed that heterozygous knockout animals spent considerably more time rearing across all three days and the overall object exploration time was not considerably different across groups (Figure A.1).

On the third day of testing, two objects from day 2 swapped places. These objects were in positions 2 and 4. A difference score was used to ascertain whether the rats spent more time with these positions of interest on the third day than the second day, indicating that they noticed the difference in the positions of the objects. Results of linear regression models

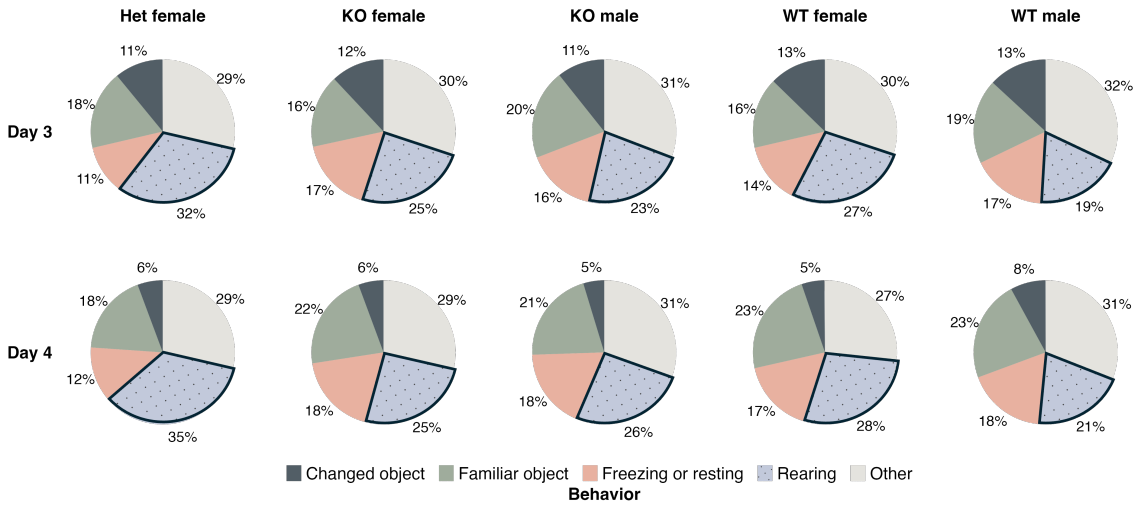


Figure A.1: *Behavior of rats on days 3 and 4. As indicated by the outlined slices in the figure, rearing behavior was greater among heterozygous females throughout the test (estimate = 0.27, estimated error = 0.10, 95% credible interval = 0.08 – 0.45). Total exploration time was not considerably different among all genotype-sex groups. KO female = homozygous knockout female, KO male = hemizygous knockout male, WT = wildtype.*

showed that while none of the groups spent more time in position 2 on day 3 compared to day 2, homozygous females (b=18.787, SE=7.882, t=2.38, p=.021), wildtype females (b=19.847, SE=7.882, t=2.52, p=.015), and wildtype males (b=18.377, SE=7.676, t=2.39, p=.021) spent more time in position 4 on day 3, compared to day 2 (Figure A.2). This data suggests that wild type rats and homozygous females noticed a change in the environment had occurred between day 2 and day 3.

On the final day of testing, the object in position 1 was replaced by a novel object. A linear regression model was run using a different score to determine whether the rats spent more time in position 1 on the final day compared to the second day of testing. Tukey-adjusted pairwise comparisons showed that wildtype male rats spent significantly more time in position 1 on the final day compared to the heterozygous knockouts (p=.001), hemizygous males (p=.033), and wildtype females (p=.002), but not homozygous females. Homozygous females also spent more time in position 1 on the final day. However, it did not survive Tukey-adjusted pairwise comparisons (Figure A.3).

In addition to tracking time spent with objects, the algorithms allowed tracking of freezing/ resting, and rearing behavior. Heterozygous females spent significantly more time rear-

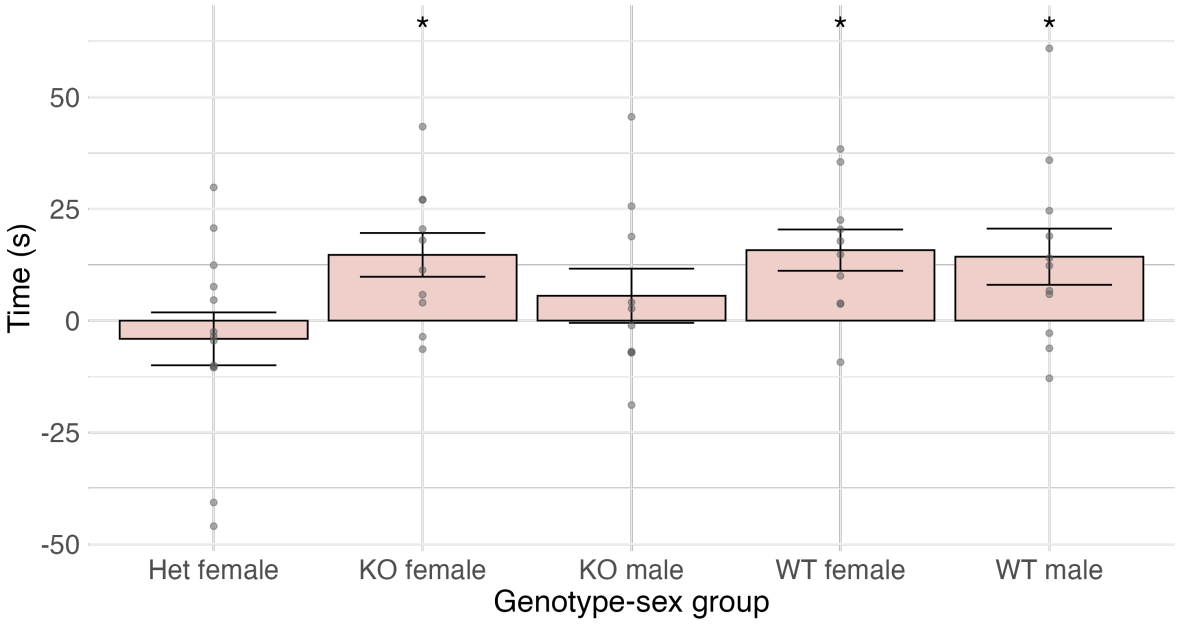


Figure A.2: *Difference in time spent in position 4 on day 3 and day 2. Homozygous females (b=18.787, SE=7.882, t=2.38, p=.021), wildtype females (b=19.847, SE=7.882, t=2.52, p=.015), and wildtype males (b=18.377, SE=7.676, t=2.39, p=.021) spent more time in position 4 on day 3 than on day 2, indicating that they noticed the change in environment. KO female = homozygous knockout female, KO male = hemizygous knockout male, WT = wildtype.*

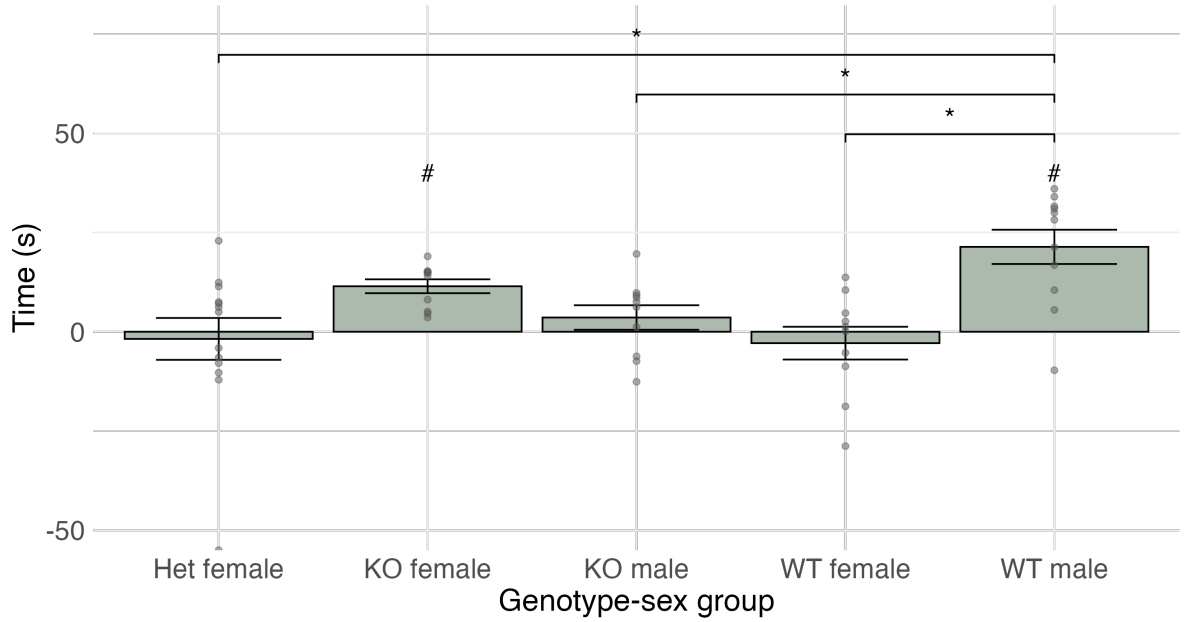


Figure A.3: *Difference in time spent in position 1 on day 4 and day 2. Wildtype males (b=23.182, SE=5.567, t=4.16, p¡.001) and homozygous females (b=13.260, SE=5.716, t=2.32, p=.025) spent more time in position 1 on day 4 than on day 2, indicating that they noticed the novel object. Pairwise comparisons show that the wildtype males spent significantly more time in position 1 on the final day compared to heterozygous knockouts (p=.001), hemizygous males (p=.033), and wildtype females (p=.002). KO female = homozygous knockout female, KO male = hemizygous knockout male, WT = wildtype, * = significant Tukey-adjusted pairwise comparison, # = significant main effect.*

ing than other rats. While wildtype males spent less time rearing than the other groups across all three days, the difference did not reach statistical significance. Homozygous females and hemizygous males had similar rearing time across all three days of testing. Heterozygous females spent less time freezing and resting than the other four groups. However, this difference was not statistically significant.

Overall, the results from this experiment show that heterozygous females spent more time rearing. Wildtype male rats, homozygous and wildtype female rats were able to recognize the changes made to the position of objects on day 3. However, when the novel object was introduced on day 4 only wildtype males and homozygous females recognized this change. Prior studies which included only male wildtype and *FMR1* knockout rodents found impairment in object recognition and object location memory in the knockouts Borreca et al. (2023); Ntoulas et al. (2024). The current results replicate those findings, where hemizygous males failed to recognize that two objects switched places on day 3, and that a novel object was introduced on day 4. Sex differences in object recognition and object location memory among wildtype rats is inconsistent in literature, with most studies reporting no sex differences, while some studies report better object recognition memory in females and better object location memory in males Becegato and Silva (2022). Here, wildtype females were not completely impaired as they noticed the switched objects and homozygous females were not impaired at all. It is possible that heterozygous females were impaired due to their fixation on rearing behavior. A future study should be carried out over more days with more changes in the arena to better capture differences in object recognition memory and object location memory.