TRACE: End-to-end temporal inference and annotation of animal behaviors from video

Kunming Shi [1,2,#], Guang-Wei Zhang [1,3,*,#], Zike Wang[1], Sonia K. Zhang [1,4],

Huizhong W. Tao [1,5], Li I. Zhang [1,5,*]

[1]Center for Neural Circuits and Sensory Processing Disorders, Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033;

[2]Graduate program in Biomedical and Biological Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033;

[3]Department of Neuroscience and Anatomy, School of Medicine, Virginia Commonwealth University, Richmond, VA 23298;

[4]Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089;

[5]Department of Physiology and Neuroscience, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.

[#] These authors contributed equally to this work

*Correspondence should be addressed to G. W. Zhang (guangwei.zhang@vcuhealth.org), or L. I. Zhang (liizhang@usc.edu)

Abstract

Quantitative analysis of animal behavior is fundamental to neuroscience and ethology but remains constrained by the scalability, subjectivity, and limited reproducibility of manual annotation. Most automated approaches infer behavior through predefined intermediate representations such as pose trajectories, which require task-specific design choices and often omit contextual visual information essential for behavioral interpretation. Here we introduce TRACE (Temporal Recognition of Animal Behaviors Captured from Video), an end-to-end method with a graphical user interface for detecting and annotating animal behavior from raw video. TRACE leverages a transformer-based video encoder pretrained via self-supervised learning to extract hierarchical temporal features, combined with multi-scale temporal modeling to capture behaviors spanning diverse timescales. The method jointly predicts behavioral identity and temporal boundaries from continuous video recordings with high-speed inference. Across multiple behavioral datasets spanning different species and experimental contexts, TRACE demonstrates robust and generalizable performance, enabling scalable, context-aware analysis of animal behavior directly from video.

Introduction

Animal behavior unfolds as temporally structured actions embedded within continuous visual streams, where both movement dynamics and surrounding context contribute to behavioral identity[1–4]. Quantitative analysis of such behavior requires identifying and localizing discrete behavioral episodes over time, a process that remains constrained by the scalability, subjectivity, and limited reproducibility of manual annotation[1–6]. Despite increasing availability of large-scale behavioral video recordings, scalable automated methods for behavioral annotation remain limited.

Most existing automated approaches infer behavior through predefined intermediate representations, most commonly pose trajectories obtained from pose-estimation pipelines[7–9]. Pose-based methods have substantially advanced behavioral quantification, providing interpretable and precise descriptions of movement across diverse experimental contexts (DeepLabCut[7], LEAP[10], SLEAP[8], DANNCE[11]). However, inferring behavioral states from pose trajectories typically requires a subsequent analytical step applied to keypoint data (MoSeq[12], MARS[13], SimBA[14]). Moreover, keypoint-based representations may not preserve contextual visual information, including posture-dependent animal's visual appearance and environmental cues, which also contribute to behavioral identity[13,15,16]. As behavioral recordings scale to longer and more continuous formats, it becomes increasingly challenging to apply these multi-stage pipelines efficiently.

Recent advances in video representation learning offer a complementary alternative. Transformer-based video encoders[17] pretrained via large-scale self-supervised learning capture hierarchical spatiotemporal representations directly from raw video, integrating motion, appearance, and context over extended temporal windows[18–20]. In human action recognition, such models enable accurate action classification[19,21] and temporal detection with minimal task-specific supervision (VideoMAE[22,23]). Yet recognizing animal behavior presents distinct challenges relative to human action recognition: behavioral datasets are considerably smaller[24,25], behavioral bouts span highly variable

durations, and recordings are typically continuous rather than pre-segmented into discrete clips[1,4]. Despite the recent advances of action recognition[19,21], direct end-to-end spatiotemporal inference of animal behavior from continuous video has not been widely adopted[15], in part because existing temporal detection frameworks have not been adapted to handle the sparsity, imbalance, and variable timescales characteristic of ethological behavior episodes.

Here we introduce TRACE, an end-to-end method for temporal inference of animal behaviors from raw video. TRACE combines a self-supervised pretrained transformer-based video encoder with multi-scale temporal modeling to jointly infer behavioral identity and temporal boundaries from continuous recordings. By eliminating reliance on predefined intermediate representations, TRACE provides a scalable and general method for discovering, annotating, and quantifying animal behavior directly from video. Across diverse laboratory and naturalistic datasets, TRACE detects temporally structured behaviors from continuous video recordings and supports accurate localization of behavioral episodes.

Results

TRACE performs end-to-end temporal inference of animal behaviors from continuous video

We developed TRACE as an end-to-end method for inferring temporally structured animal behaviors directly from video recordings. The overall design of TRACE and the formulation of the temporal behavior detection task are illustrated in **Figure 1**. Video recordings are annotated by human experts with discrete behavioral identities and corresponding start–stop temporal boundaries, producing temporally localized behavioral segments spanning a wide range of durations, using our customized graphic user interface (see **Methods**). These annotated segments serve as training data for learning a direct mapping between raw video input and temporally resolved behavioral outputs.

TRACE processes raw video using a transformer-based video encoder[17] that operates on video volumes. Video frames are divided into fixed-length temporal chunks (**Fig. 1b**). The model processes chunks of video frame and uses self-attention so that each frame feature representation can incorporate information from other frames across time. This enables the model to capture temporal dependencies, including action sequences, transitions, and evolving context. So instead of treating each frame independently, the model understands how frames are connected across time. To accommodate the variable temporal scales of animal behaviors, TRACE projects these frame-wise features into a multi-scale temporal feature pyramid (**Fig. 1c**), enabling detection of short-and long-duration behaviors.

Temporal detection is performed using a trident detection head[21] composed of three parallel prediction branches (**Fig. 1d**). For each candidate behavioral instance, TRACE jointly predicts behavioral identity, temporal center location, and start–stop boundaries, allowing localization of discrete behavioral bouts within recordings (**Fig. 1e**). All components of the method—including the video encoder, projection module, and detection head—are optimized jointly using classification and temporal regression losses (**Fig. 1f**). After training, TRACE performs inference directly on raw video recordings to generate temporally localized behavioral predictions without human annotation (**Fig. 1g**).

Automated temporal detection of spontaneous mouse behaviors

We first evaluated TRACE with ViT-Large variant on freely behaving mice annotated with four spontaneous behaviors: self-grooming, rearing, drinking, and eating (**Fig. 2a**). These behaviors span diverse movement dynamics, postural configurations, and temporal durations, providing a stringent test of TRACE's ability to perform temporally resolved behavioral inference. Raster plots of human annotations across training clips illustrate the temporal structure and variability of behavioral bouts within the recordings (**Fig. 2b**), as well as substantial background periods without labeled behavior ("no-label").

Across the clips used for training (80% of the entire dataset), the relative proportion of time spent in each behavior varied considerably, with a substantial fraction of frames remaining unlabeled (**Fig. 2c**), reflecting the natural imbalance and sparsity typical of spontaneous behavior. Despite these challenges, TRACE converged reliably during training, with training loss (total, classification, and regression) decreasing steadily over training epochs (**Fig. 2d**). Behavior-specific learning curves demonstrated robust performance across all four behaviors, with precision improving consistently during training (**Fig. 2e**).

We further assessed the robustness of TRACE to variations in video acquisition and training conditions. Reducing the spatial resolution of video frames or frame rate resulted in modest decrease in performance, as measured by the mean Average Precision (mAP) (**Fig. 2f** & **g**, orange), within the range of conditions tested. Reducing the amount of training data as low as 4% led to only limited reduction in performance under the original acquisition settings (**Fig. 2h**, orange). In addition, TRACE with ViT-Small variant achieved comparable performance under the original video parameters (1080 pixels, 30 frames per second) in this dataset (**Fig. 2f-h**, blue), suggesting that model capacity can be reduced without substantial loss of accuracy under these conditions.

After training, TRACE (ViT-Large) was applied to the remaining 20% of the dataset to predict behaviors on held-out videos. TRACE-predicted behavioral bouts closely matched human annotations in video segments (**Fig. 2i**), with high precisions and recalls across behavioral classes and effective separation from background activity (**Fig. 2j**, **k**). In addition, accuracy per video clip in both ViT-Small and ViT-Large variants reached approximately 95% (**Fig. 2l**), indicating accurate temporal localization of behaviors.Confusion-matrix (CM) analysis revealed strong agreements between model predictions and human annotations. Inference speed analysis further showed that TRACE achieved high-throughput processing, reaching over 12,500 frames per second (FPS) depending on the batch size (**Fig. 2m**).

Through automated behavior detection, TRACE enabled quantitative comparison of behavioral patterns between experimental groups. Analysis of video recordings from 5×FAD Alzheimer's disease

(AD) model mice[26]and wild-type controls in open-field conditions revealed significant differences in behavioral temporal patterns (**Fig. 2n**). Specifically, AD mice exhibited increased rearing ($p = 0.038$) and reduced self-grooming ($p = 1.6 \times 10^{-4}$) compared with wild-type animals, whereas drinking and eating behaviors did not differ significantly between the two groups (**Fig. 2o**–**r**). These results demonstrate that TRACE can be used to extract biologically meaningful behavioral phenotypes from long-duration recordings.

Temporal detection of mouse social behaviors in the CalMS21 benchmark

To evaluate TRACE (ViT-Large variant) on a standardized behavioral benchmark, we applied the model to the Caltech Mouse Social Interactions (CalMS21) dataset[27]. This dataset contains annotated recordings of resident–intruder interactions with three primary behavioral classes—attack, investigation, and mounting—along with background frames lacking behavioral labels ("no label") (**Fig. 3a**). Raster plots of annotated clips illustrate the temporal organization of social behaviors within the recordings (**Fig. 3b**). This dataset is highly imbalanced, with most frames labeled as background or investigation (**Fig. 3c**, **d**). Despite this imbalance, TRACE converged reliably during training, as reflected by steadily decreasing training losses (**Fig. 3e**).

On the CalMS21 benchmark, TRACE achieved a mAP of 94.5%, outperforming the baseline method (88.9%)[27], the competition Top-1 model (91.4%)[27], and Google's VideoPrism model (91.5%)[28]. Training curves showed rapid improvement in mAP during early epochs before stabilizing at this performance level (**Fig. 3f**). Per-class average precision (AP) curves demonstrated stable learning for attack, investigation, and mounting behaviors (**Fig. 3g**–**i**). TRACE predictions closely matched human annotations across continuous recordings, as shown by ethograms of example clips (**Fig. 3j**). Precision and recall confusion matrices further demonstrated strong agreements between predicted and annotated behaviors (**Fig. 3k**). Per-video accuracy analysis showed that the majority of videos achieved accuracies

above 88%, with a median exceeding 90% (**Fig. 3l**), while using ViT-Small only slightly reduced the accuracy. Inference speed analysis showed highly efficient processing across model variants (**Fig. 3m**), suggesting reliable and efficient discrimination among social interaction behaviors.

Cross-species behavioral detection in Drosophila and chimpanzee datasets

To evaluate the generality of TRACE (ViT-Large) across species and behavioral contexts, we applied the model to datasets capturing behaviors in *Drosophila melanogaster* and wild chimpanzees. In the fruit fly dataset[29], TRACE detected courtship behaviors including circling, copulation, and wing extension (**Fig. 4a**). Raster plots of annotated training clips illustrate the temporal distribution of these behaviors (**Fig. 4b**), with copulation representing the majority of labeled frames (**Fig. 4c**,**d**). Training curves showed stable convergence of classification and regression losses (**Fig. 4e**), and mean average precision reached 86.3% at convergence (**Fig. 4f**). Per-behavior training curves indicated high detection accuracy for copulation and strong performance for circling and wing-extension behaviors (**Fig. 4g**–**i**). Visual comparison of predicted and annotated timelines confirmed accurate temporal localization of behavioral bouts (**Fig. 4j**), and confusion-matrix analysis demonstrated high precision and recall across behavioral classes (**Fig. 4k**,**l**), with high-speed processing for both ViT-Small and -Large variants (**Fig. 4m**).

We next evaluated TRACE on camera-trap recordings of wild chimpanzees from the PanAf dataset[30] (**Fig. 4n**). These recordings contain diverse locomotor and postural behaviors under natural environmental conditions. TRACE successfully detected multiple behavior categories across recordings (**Fig. 4n**), with training losses converging steadily during optimization (**Fig. 4o**). Per-class average precision analysis revealed high detection accuracy for common behaviors such as sitting, walking, and standing, while rarer behaviors such as hanging showed much lower performance (**Fig. 4p,q**), and inference speed analysis indicated similarly efficient processing for the PanAf dataset (**Fig. 4r**). The

lower performance for certain behaviors likely reflects class imbalance and the clip-level multi-label annotation scheme of the PanAf dataset, in which co-occurring behaviors are not assigned to individual animals. Nevertheless, the successful detection of common behaviors across naturalistic camera-trap recordings demonstrates the generalizability of TRACE to wildlife datasets without species-specific adaptation.

Together, these results demonstrate that TRACE can perform temporal detection of animal behaviors across diverse species, behavioral repertoires, and recording environments without requiring species-specific pose representations.

Discussion

In this study, we developed TRACE, an end-to-end method for temporal inference of animal behaviors. Across multiple datasets spanning laboratory and naturalistic conditions, including mouse social interactions, fly behavior, chimpanzee camera-trap footage, and a mouse model of Alzheimer's disease, TRACE consistently detected temporally structured behaviors with localized start–stop boundaries. On the CalMS21[27] social interaction benchmark, TRACE achieved higher mean average precision than previously reported baseline and benchmark models[27,28], indicating that this approach can provide accurate behavioral detection without requiring intermediate pose representations. By jointly predicting behavioral identity and temporal boundaries, TRACE enables structured quantification of behavioral patterns in continuous recordings and supports scalable behavioral analysis across diverse experimental contexts. Unlike pose-based approaches[7,8] that infer behavior through intermediate keypoint representations, TRACE learns spatiotemporal representations that integrate motion and visual context, while remaining compatible with existing pose-based workflows.

Despite the growing interest in inferring behavior directly from video across the field[28,31,32], direct inference of behavior has remained challenging due to the high dimensionality of video data and the

limited availability of labeled behavioral datasets[1,2]. At a technical level, TRACE builds on advances in self-supervised video representation learning[17,20], leveraging pretrained transformer-based encoders[17]to support efficient adaptation to behavioral tasks. These hierarchical spatiotemporal representations can be fine-tuned with comparatively modest annotated datasets (tens of clips), enabling behavioral inference without task-specific handcrafted features. By integrating multi-scale temporal modeling and framing behavior recognition as a temporal detection problem[21] rather than frame-wise classification alone, TRACE can explicitly model the event-based structure of behavioral annotations. This framework provides a standalone approach for scalable behavioral annotation, with particular advantages in settings where contextual visual information contributes to behavioral identity. In experimental contexts where fine-grained kinematic measurements are additionally required, TRACE can be integrated with pose-estimation tools[7,8], combining the strengths of both approaches. In addition, TRACE supports high-throughput analysis through efficient inference, enabling rapid processing of large-scale behavioral video datasets.

Several limitations warrant consideration. TRACE relies on supervised annotations, and its performance depends on the quality and consistency of behavioral labels in the training data. The method does not independently discover behavioral categories and remains constrained by predefined class definitions. While video-based inference captures contextual information beyond pose, it does not provide explicit kinematic measurements or identity tracking and may be most effective when integrated with complementary tracking systems[7,8]. In addition, the transformer-based video encoders used in TRACE require greater computational resources than lightweight pose-estimation pipelines; however, both training and inference can be performed on high-memory laboratory workstations equipped with modern GPUs.

Taken together, TRACE illustrates how end-to-end spatiotemporal modeling can complement pose-based approaches for behavioral quantification. Future extensions of TRACE may incorporate

alternative video encoders[33–35] as advances in video representation learning continue. By integrating contextual visual information with explicit temporal detection, TRACE provides a method for structured temporal inference of animal behavior while remaining compatible with existing kinematic and tracking pipelines.

Acknowledgments

This work was supported by grants from the U.S. National Institutes of Health (DC008983, DC020887, NS132912 to L.I.Z. and EY019049, AG089756 to H.W.T.), and Alzheimer Foundation Fellowship (AARF-23-1148428) to G.W.Z. High performance computing resources provided by the High Performance Research Computing (HPRC) core facility at Virginia Commonwealth University (https://hprc.vcu.edu) were used for conducting the research reported in this work. The Center for Advanced Research Computing (CARC) at the University of Southern California (https://carc.usc.edu) also provided computing resources for this study.

Author Contributions

L.I.Z. and G.W.Z. conceived the study and supervised the project. K.S. and G.W.Z. adapted, implemented, and tested the TRACE model, designed and performed behavioral experiments, and conducted data analysis. S.Z. and Z.W. contributed to data analysis. L.I.Z., H.W.T., and G.W.Z. wrote the manuscript.

Methods

Animals

All experimental procedures in this study have been approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Southern California. Male and female adult mice, C57BL/6J and 5xFAD, were used in this study. Mice were housed at 18–23 °C with 40–60% humidity in a 12-h light-dark cycle (6AM-6PM) with ad libitum access to food and water.

Datasets.

We evaluated TRACE across four datasets spanning diverse species, recording conditions, and levels of behavioral organization. All datasets provide raw video paired with temporally annotated behavioral episodes, enabling temporal action detection with explicit start and end time boundaries. Training and evaluation details for each dataset are described below. Across all datasets, behavioral events were treated as non-overlapping and each frame was assigned at most one behavior label during annotation.

Single Mouse Behavior Dataset.

To assess TRACE on spontaneous single-animal behaviors, we compiled a dataset of self-recorded side-view videos of adult wild-type C57BL/6J mice freely behaving in their home cage. Videos were captured from a lateral perspective at 30 frames per second (fps) under standard housing conditions. Behavioral episodes were manually annotated by trained observers using TRACE event-logging software, covering four ethologically defined categories: rearing, self-grooming, drinking, and eating. These behaviors span a wide range of durations and display profiles, from brief, repetitive self-grooming to extended eating and drinking. We divided the dataset into training and validation splits. The dataset comprises 10 source videos recorded across multiple sessions, split into 506 clips (400 training, 106 validation), yielding a total of 1,414 annotated behavioral events and evaluated model performance using mean Average Precision (mAP).

Caltech Mouse Social Interaction (CalMS21).

The CalMS21 dataset[27] captures social interactions between pairs of freely behaving laboratory mice during a resident–intruder assay, recorded from a top-view camera at 30 fps. The dataset contains frame-level behavioral annotations across three ethologically defined categories: attack, investigation, and mount, reflecting distinct patterns of agonistic and affiliative interactions. We used the standard train–validation split from the original challenge. Unlabeled frames were treated as a separate "other" background class to enable the model to suppress false detections during inactive periods. Model performance was evaluated using mAP.

Caltech Fly Courtship Dataset.

This dataset contains videos of pairs of Drosophila melanogaster engaged in courtship interactions[29]. Videos were recorded from a dorsal view under controlled laboratory conditions at 30 fps, with expert-annotated temporal boundaries for three courtship-specific behaviors: wing extension (the male's courtship "song" display), circling, and copulation. These behaviors are characterized by brief, rapidly alternating bouts and require precise temporal localization. We used the standard train–validation split and evaluated model performance using mAP.

Pan African Programme Primate Behavior (PanAf500).

The PanAf500 dataset[30] is a subset of the PanAf20K collection, comprising 500 camera-trap video clips of wild chimpanzees (Pan troglodytes) and gorillas recorded across 18 field sites in tropical Africa as part of the Pan African Programme: The Cultured Chimpanzee. The videos capture a wide range of locomotor, postural, and social behaviors under diverse lighting and environmental conditions. Behavioral annotations are provided at the individual level for each clip. The annotated behavior categories include walking, standing, sitting, climbing up, climbing down, hanging, running, sitting on back, and camera interaction. We followed the original train/validation/test split but merged the validation and test sets into a single, larger validation set. Per-individual annotations were pooled to construct video-level multi-label targets. Model performance was evaluated using mAP.

Model Architecture

TRACE is built upon TriDet[21], a one-stage temporal action detection framework that jointly predicts action class probabilities and precise temporal boundaries from continuous video features. The full pipeline consists of three sequential components: a spatiotemporal feature backbone, a multi-scale temporal feature pyramid, and a dense prediction head.

Backbone: Video-Adapted Vision Transformer. Spatiotemporal features are extracted using a Vision Transformer (ViT) augmented with parameter-efficient adapter modules (VisionTransformerAdapter). Vision Transformers provide strong capacity for long-range spatiotemporal modeling, making them well suited for capturing extended behavioral context across video sequences. Across all datasets, we evaluated two backbone variants for TRACE: ViT-Large variant (embed_dim = 1024, 24 transformer layers, 16 attention heads; ~304 million parameters), and ViT-Small variant (embed_dim = 384, 12 transformer layers, 6 attention heads; ~22 million parameters). Both were pretrained via VideoMAE masked autoencoding on the Kinetics-400 dataset. We additionally evaluated a The backbone uses a patch size of 16×16 pixels and processes clips of 16 frames per temporal token. Input frames were resized to 144 × 144 pixels (224 × 224 for PanAf500) before patch embedding. During fine-tuning, backbone weights were frozen to reduce computational cost and mitigate overfitting on relatively small behavioral datasets, and only lightweight adapter modules (~ 7% of total parameters) inserted at each transformer layer were updated. This parameter-efficient design substantially reduces computational cost while preserving the rich spatiotemporal representations learned during pretraining. Long video sequences were processed by dividing each video into non-overlapping 16-frame micro-clips, each independently encoded, then concatenated and temporally interpolated to produce a dense feature sequence with one feature vector per frame.

Temporal Feature Pyramid (TriDetProj and FPNIdentity). The frame-level feature sequence extracted by the backbone is passed into the TriDet projection module (TriDetProj), which builds a hierarchical

multi-scale temporal representation. The projection module consists of two embedding convolutional layers followed by a stem of two Scalable-Granularity Perception (SGP) blocks and a five-level branch with progressive 2× temporal downsampling, producing feature maps at six temporal scales with strides [1, 2, 4, 8, 16, 32] relative to the input. Each SGP block aggregates local temporal context using depthwise convolutions with a kernel size of 3 and projects features through an MLP with hidden dimension 768. Absolute positional encodings are applied to the input sequence. The resulting multi-scale feature pyramid is passed through a Feature Pyramid Network identity connection (FPNIdentity), maintaining 512 channels at all six scales.

Detection Head: TriDetHead. Temporal action predictions are generated by the TriDetHead, a dense anchor-free detection head that operates independently on each scale of the feature pyramid. At each temporal location and scale, the head predicts: (1) class probabilities across all behavior categories, and (2) temporal boundary offsets encoding the distance to the start and end of the behavioral episode. Boundary regression employs a trident representation that combines a center-offset prediction with learned probability distributions over 16 discrete bins for start and end boundary positions, enabling sub-frame boundary precision. Classification confidence and boundary quality are jointly used during inference to rank proposals. Candidate proposals are filtered and merged using Gaussian Soft-NMS ($\sigma = 0.5$) with a voting threshold of 0.7, retaining up to 2,000 segments per video.

Training

All models were trained end-to-end using the AdamW optimizer with a weight decay of 0.025. Backbone adapter parameters used a separate, higher learning rate than the remaining model components to account for the difference in parameter scale and convergence dynamics. Training followed a linear warmup schedule for the first 5 epochs, after which the learning rate decayed according to a cosine annealing schedule.

During training, video clips were randomly truncated to the target window size and augmented with random resized cropping (scale randomly sampled from [0.9, 1.0] of the original area), horizontal

flipping (p = 0.5), standard image augmentations (brightness, contrast, sharpness), and color jitter. For validation and testing, clips were processed using a sliding-window protocol of 768 frames (368 for PanAf500) corresponding to approximately 25.6 sec (12.3 sec for PanAf500) at 30 fps, with 25% and 50% overlap respectively, and crops were taken from the center of each window. Predictions from overlapping windows were merged using Gaussian Soft-NMS. All frames were normalized using ImageNet statistics (mean = [123.675, 116.28, 103.53], std = [58.395, 57.12, 57.375]). Training used automatic mixed precision (AMP) with FP16 gradient compression and gradient clipping (max norm = 1). Exponential moving average (EMA) of model weights was maintained throughout training for stable evaluation. Training was performed on a single NVIDIA H100 GPU. Training required approximately 20 hours (100 epochs) for the single mouse dataset, 34 hours (300 epochs) for CalMS21, and 67 hours (300 epochs) for the Caltech fly courtship dataset, with a batch size of 2 (1 for CalMS21). Inference was benchmarked on a NVIDIA H100 GPU. In addition, both the ViT-Small and ViT-Large variants were evaluated on a laboratory server equipped with an NVIDIA RTX 4060 Ti GPU, where training and inference could be performed.

Loss Functions

TRACE is trained with a combination of three loss terms. Classification is supervised by Focal Loss, which down-weights easy background examples and focuses learning on rare, ambiguous behavioral events—particularly important for datasets with pronounced class imbalance between background and active behavior periods. Temporal locations that do not overlap with any annotated behavioral episode are treated as background; the model learns to suppress these through the Focal Loss weighting, which assigns lower loss to easily classified background frames and higher loss to ambiguous or rare positive detections. Temporal boundary regression is supervised by Distance-IoU (DIOU) Loss, which penalizes both the distance between predicted and ground-truth segment centers and the overlap deficit, providing stable gradients for precise boundary localization. In addition, an IoU quality branch is trained with Generalized IoU (GIOU) Loss and weighted by IoU quality raised to a power of 0.2, enabling the model

to jointly learn proposal confidence and temporal overlap. The total loss normalizer is maintained at 100 using an exponential moving average (momentum = 0.9) to stabilize gradient magnitudes across training steps.

Evaluation

Model performance was evaluated using mAP at multiple IoU thresholds: 0.3, 0.4, 0.5, 0.6, and 0.7. The mAP across these thresholds served as the primary summary metric. This evaluation protocol, adapted from temporal action detection benchmarks, quantifies not only whether the model detects the correct behavior category, but also how accurately it localizes the temporal boundaries of each behavioral episode. Higher IoU thresholds require tighter temporal boundary alignment, reflecting the stringency of the evaluation. The 'other' background class was excluded from all reported metrics. For all datasets, predictions were generated using the sliding-window test protocol with 50% overlap, and the final ethogram was assembled by merging detections from all windows using Soft-NMS.

Statistical Analysis

To compare behavioral profiles between wild-type (WT, n = 3) and Alzheimer's disease model (AD, n = 3) mice, TRACE-generated ethograms were used to quantify the percentage of recording time spent in each behavioral category (rearing, self-grooming, drinking, and eating) per video. Group differences were assessed using two-sided Mann–Whitney U tests (scipy.stats.mannwhitneyu) for each behavior independently. Significance levels were defined as * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$. No corrections for multiple comparisons were applied. Cumulative behavior time curves were plotted as mean ± SEM across videos within each group.

Code Availability

The Graphical User Interface, source code for TRACE, including training and inference scripts, pretrained model weights, and example configurations for all datasets, are publicly available on GitHub: https://github.com/KunmingS/TRACE.

References

1. Anderson, D. J. & Perona, P. Toward a science of computational ethology. *Neuron* **84**, 18–31 (2014).
2. Datta, S. R., Anderson, D. J., Branson, K., Perona, P. & Leifer, A. Computational Neuroethology: A Call to Action. *Neuron* **104**, 11–24 (2019).
3. Mathis, M. W. & Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Curr Opin Neurobiol* **60**, 1–11 (2020).
4. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. *Nat Neurosci* **23**, 1537–1549 (2020).
5. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J R Soc Interface* **11**, 20140672 (2014).
6. Kafkafi, N. *et al.* Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neurosci Biobehav Rev* **87**, 218–232 (2018).
7. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).
8. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat Methods* **19**, 486–495 (2022).
9. Willmore, L., Cameron, C., Yang, J., Witten, I. B. & Falkner, A. L. Behavioural and dopaminergic signatures of resilience. *Nature* **611**, 124–132 (2022).
10. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat Methods* **16**, 117–125 (2019).
11. Dunn, T. W. *et al.* Geometric deep learning enables 3D kinematic profiling across species and environments. *Nat Methods* **18**, 564–573 (2021).
12. Weinreb, C. *et al.* Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nat Methods* **21**, 1329–1339 (2024).

13. Segalin, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *Elife* **10**, e63720 (2021).
14. Goodwin, N. L. *et al.* Simple Behavioral Analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nat Neurosci* **27**, 1411–1424 (2024).
15. Bohnslav, J. P. *et al.* DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife* **10**, e63377 (2021).
16. Simonyan, K. & Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. Preprint at https://doi.org/10.48550/ARXIV.1406.2199 (2014).
17. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/ARXIV.2010.11929 (2020).
18. Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Preprint at https://doi.org/10.48550/ARXIV.2103.14030 (2021).
19. Zhang, C., Wu, J. & Li, Y. ActionFormer: Localizing Moments of Actions with Transformers. Preprint at https://doi.org/10.48550/ARXIV.2202.07925 (2022).
20. Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. Preprint at https://doi.org/10.48550/ARXIV.2304.07193 (2023).
21. Shi, D. *et al.* TriDet: Temporal Action Detection with Relative Boundary Modeling. Preprint at https://doi.org/10.48550/ARXIV.2303.07347 (2023).
22. Wang, L. *et al.* VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. Preprint at https://doi.org/10.48550/ARXIV.2303.16727 (2023).
23. Tong, Z., Song, Y., Wang, J. & Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. Preprint at https://doi.org/10.48550/ARXIV.2203.12602 (2022).
24. Ng, X. L. *et al.* Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding. Preprint at https://doi.org/10.48550/ARXIV.2204.08129 (2022).

25. Chen, J. *et al.* MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding. Preprint at https://doi.org/10.48550/ARXIV.2306.00576 (2023).

26. Oakley, H. *et al.* Intraneuronal β-Amyloid Aggregates, Neurodegeneration, and Neuron Loss in Transgenic Mice with Five Familial Alzheimer's Disease Mutations: Potential Factors in Amyloid Plaque Formation. *J. Neurosci.* **26**, 10129–10140 (2006).

27. Sun, J. J. *et al.* The Multi-Agent Behavior Dataset: Mouse Dyadic Social Interactions. Preprint at https://doi.org/10.48550/arXiv.2104.02710 (2021).

28. Zhao, L. *et al.* VideoPrism: A Foundational Visual Encoder for Video Understanding. Preprint at https://doi.org/10.48550/arXiv.2402.13217 (2025).

29. Eyjolfsdottir, E. *et al.* Detecting Social Actions of Fruit Flies. in *Computer Vision – ECCV 2014* (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) vol. 8690 772–787 (Springer International Publishing, Cham, 2014).

30. Brookes, O. *et al.* PanAf20K: A Large Video Dataset for Wild Ape Detection and Behaviour Recognition. Preprint at https://doi.org/10.48550/arXiv.2401.13554 (2024).

31. Hu, Y. *et al.* LabGym: Quantification of user-defined animal behaviors using learning-based holistic assessment. *Cell Reports Methods* **3**, 100415 (2023).

32. Lin, S. *et al.* Characterizing the structure of mouse behavior using Motion Sequencing. *Nat Protoc* **19**, 3242–3291 (2024).

33. Zhao, L., Gundavarapu, N. B. & Yuan, L. VideoPrism: A foundational visual encoder for video understanding. in *Proceedings of the International Conference on Machine Learning* 60785–60811 (2024). doi:10.48550/arXiv.2402.13217.

34. Wang, Y., Li, K. & Li, X. InternVideo2: Scaling foundation models for multimodal video understanding. in *European Conference on Computer Vision* 396–416 (2024). doi:10.48550/arXiv.2403.15377.

35. Oquab, M., Darcet, T. & Moutakanni, T. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* https://doi.org/10.48550/arXiv.2304.07193 (2024) doi:10.48550/arXiv.2304.07193.

Figure Legends

Figure 1 TRACE architecture for transformer-based behavioral temporal detection.

a, Annotated video training data. Raw videos are manually annotated with action onset and offset boundaries to generate temporal action sequences used as supervision during training. **b,** Video encoder based on a Vision Transformer (ViT). Each input video volume is divided into 48 non-overlapping temporal chunks of 16 frames, embedded, and processed through self-attention layers to produce a temporally resolved feature matrix. **c,** Multi-scale temporal pyramid. Projected features are represented at multiple temporal resolutions to capture behavioral events spanning short to long durations. **d,** Trident Detection Head. The detection head contains three parallel branches: a classification head, a center-regression head, and two boundary heads for predicting action category and temporal extent. **e,** Behavioral temporal detection scheme. For each action instance, TRACE predicts the behavioral class, temporal center, and corresponding start and stop boundaries. **f,** Training pipeline. Annotated videos are processed through the video encoder, projection module, and trident detection head in a forward pass; loss is backpropagated through all modules to optimize model parameters. **g,** Inference pipeline. After training, raw videos are processed by the trained network to generate behavioral temporal detections without manual annotation or pose estimation.

Figure 2 TRACE achieves accurate temporal detection of mouse home-cage behaviors and reveals behavioral alterations in 5×FAD mice.

a, Representative frames illustrating the four annotated home-cage behaviors: self-grooming, drinking, rearing, and eating. **b,** Visualization of the training dataset across five partitions. Each row represents one training clip, with colored segments denoting annotated behavioral bouts. **c,** Behavioral composition of the training dataset. Pie chart, per-frame proportions including self-grooming, drinking, rearing, eating, and unlabeled intervals; bar plot, relative proportions among labeled behaviors. **d,** Training loss curves across 150 epochs showing total, classification, and regression losses. **e,** Detection

performance during training. Per-class mean average precision (mAP) across epochs. **f–h,** Robustness analyses showing mAP as a function of input resolution (**f**), frame rate (**g**), and training data fraction (**h**) for small and large TRACE variants. **i,** Comparison of ground-truth and TRACE-predicted behavioral timelines across video clips, showing temporal localization of behavioral bouts. **j, k,** Precision (**j**) and recall (**k**) confusion matrices for rearing, self-grooming, drinking, eating, and background classes. **l,** Distribution of per-video classification accuracy for small and large TRACE variants. **m,** Inference speed (frames per second; FPS) as a function of batch size. **n,** Behavioral timelines across 60-min recordings from 5×FAD and wild-type (WT) mice, with each row representing one recording session. **o–r,** Cumulative time spent in rearing (**o**), self-grooming (**p**), drinking (**q**), and eating (**r**) in 5×FAD and WT mice. Shaded regions indicate mean ± s.e.m. P values were calculated using two-sided Mann–Whitney U tests.

Figure 3 TRACE accurately localizes social behaviors in the CalMS21 mouse social interaction benchmark.

a, Representative frames illustrating investigation, attack, and mount behaviors. **b,** Visualization of the training dataset across six partitions. Each row represents one training clip, with colored segments indicating annotated behavioral bouts and unlabeled intervals. **c,** Pie chart showing per-frame behavioral composition of the dataset. **d,** Relative proportions of labeled behavioral classes. **e,** Training loss curves over 300 epochs for total, classification, and regression losses. **f,** Mean average precision (mAP) during training. **g–i,** Per-class AP curves for attack (**g**), investigation (**h**), and mount (**i**). **j,** Comparison of ground-truth and TRACE-predicted behavioral timelines across video clips. **k,** Precision and recall (right) confusion matrix. **l,** Distribution of per-video classification accuracy for small and large TRACE models. **m,** Inference speed as a function of batch size for small and large TRACE variants.

Figure 4 TRACE generalizes across species, detecting courtship behaviors in Drosophila and diverse natural behaviors in chimpanzees.

a, Representative frames of drosophila courtship behaviors: copulation, circle, and wing extension. **b,** Visualization of the drosophila training dataset across eight partitions. Each row represents one training clip, with colored segments indicating circle, copulation, wing extension, and unlabeled intervals. **c,** Pie chart of per-frame behavioral composition in the drosophila dataset. **d,** Relative proportions of labeled Drosophila behaviors. **e,** Training loss curves for the drosophila dataset. **f,** Mean average precision (mAP) during training. **g–i,** Per-class AP curves for circle (**g**), copulation (**h**), and wing extension (**i**). **j,** Comparison of ground-truth and TRACE-predicted drosophila behavioral timelines across video clips. **k, l,** Precision (**k**) and recall (**l**) confusion matrices for the drosophila dataset. **m,** Inference speed (frames per second) as a function of batch size. **n,** Representative chimpanzee image and behavioral composition of the chimpanzee dataset. **o,** Training loss curves for the chimpanzee dataset. **p,** Per-class average precision for chimpanzee behaviors. **q,** Ground-truth and TRACE-predicted chimpanzee behavioral timelines. **r,** Inference speed as a function of batch size.

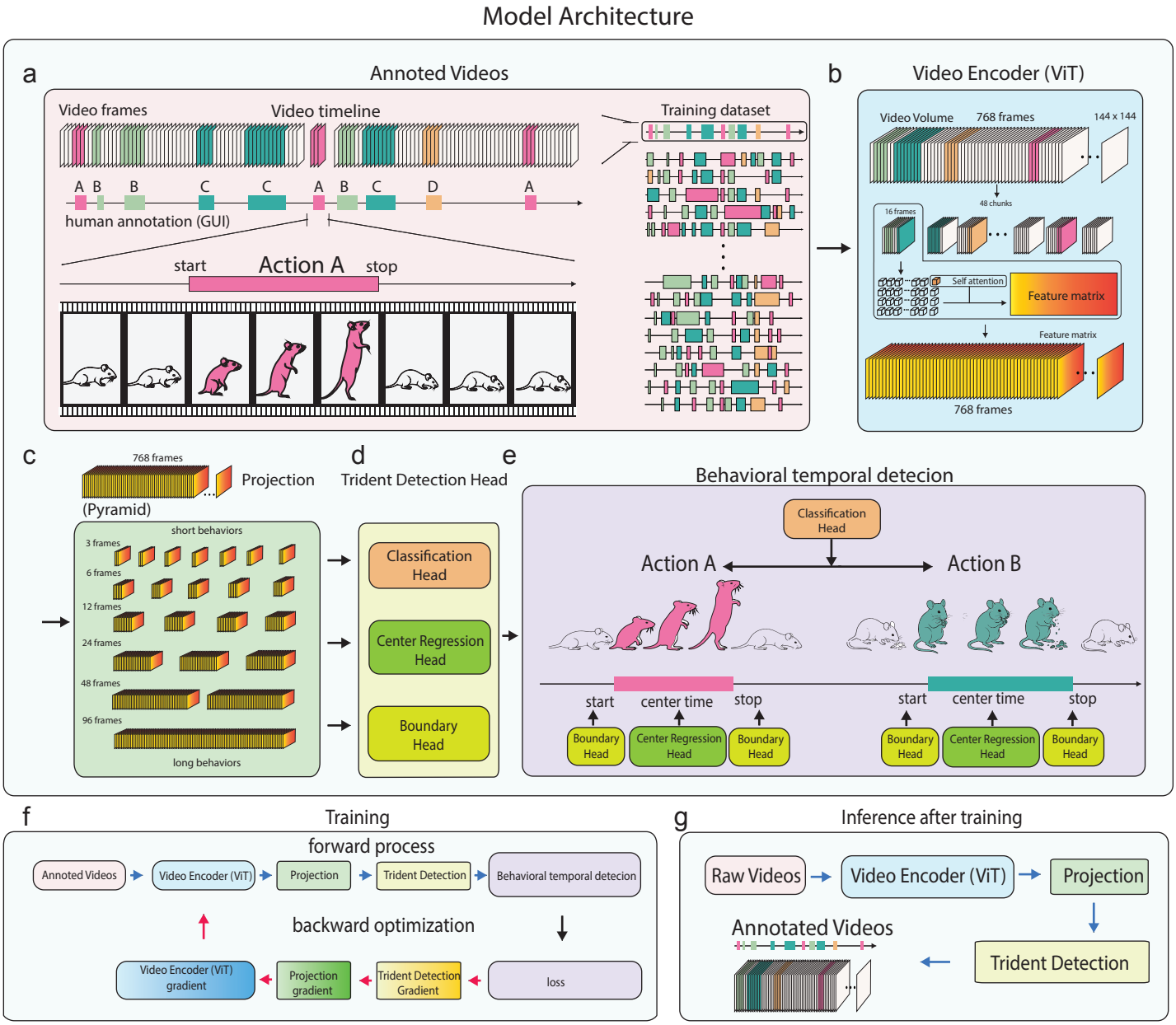

Figure 1

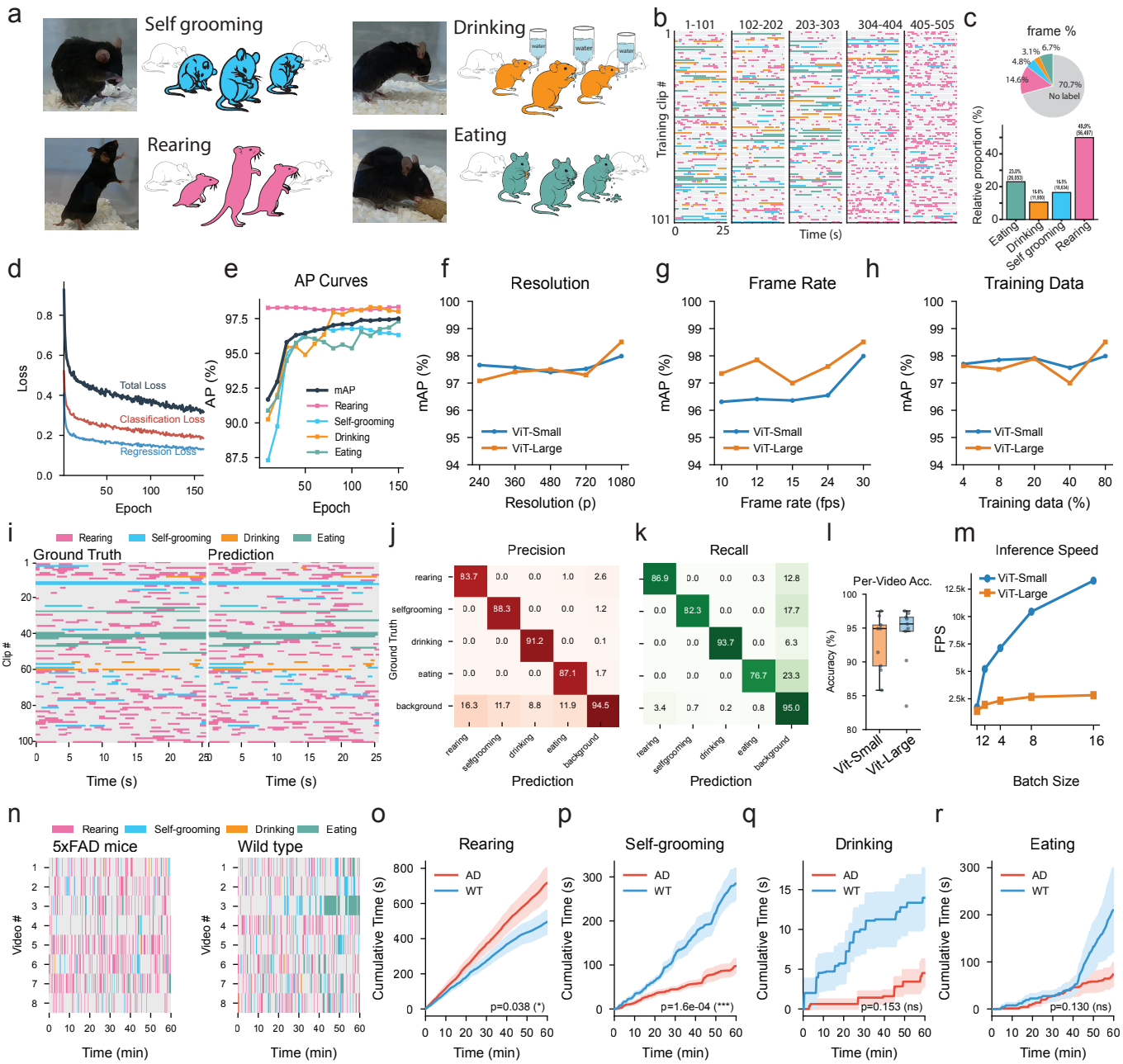

Figure 2

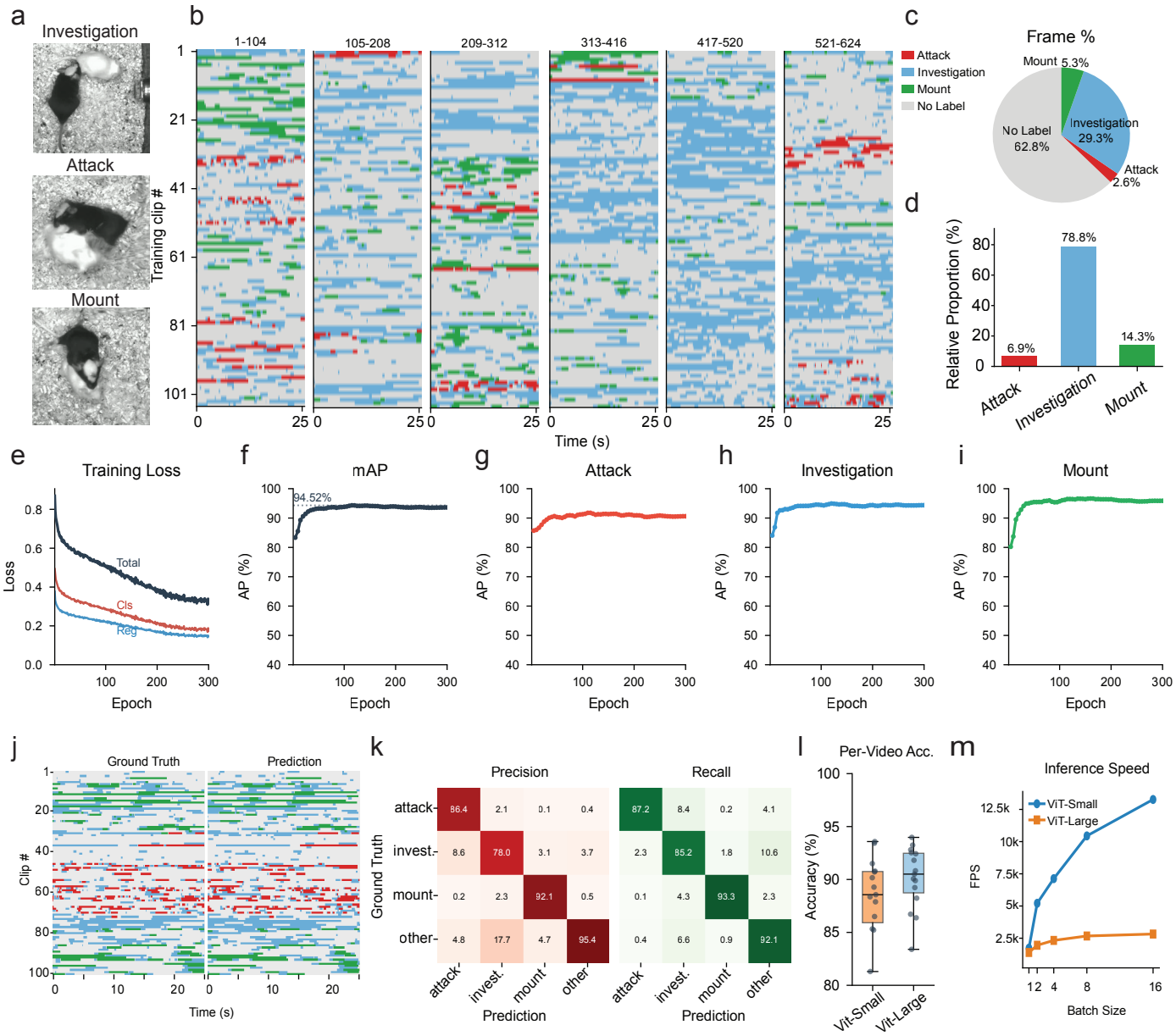

Figure 3

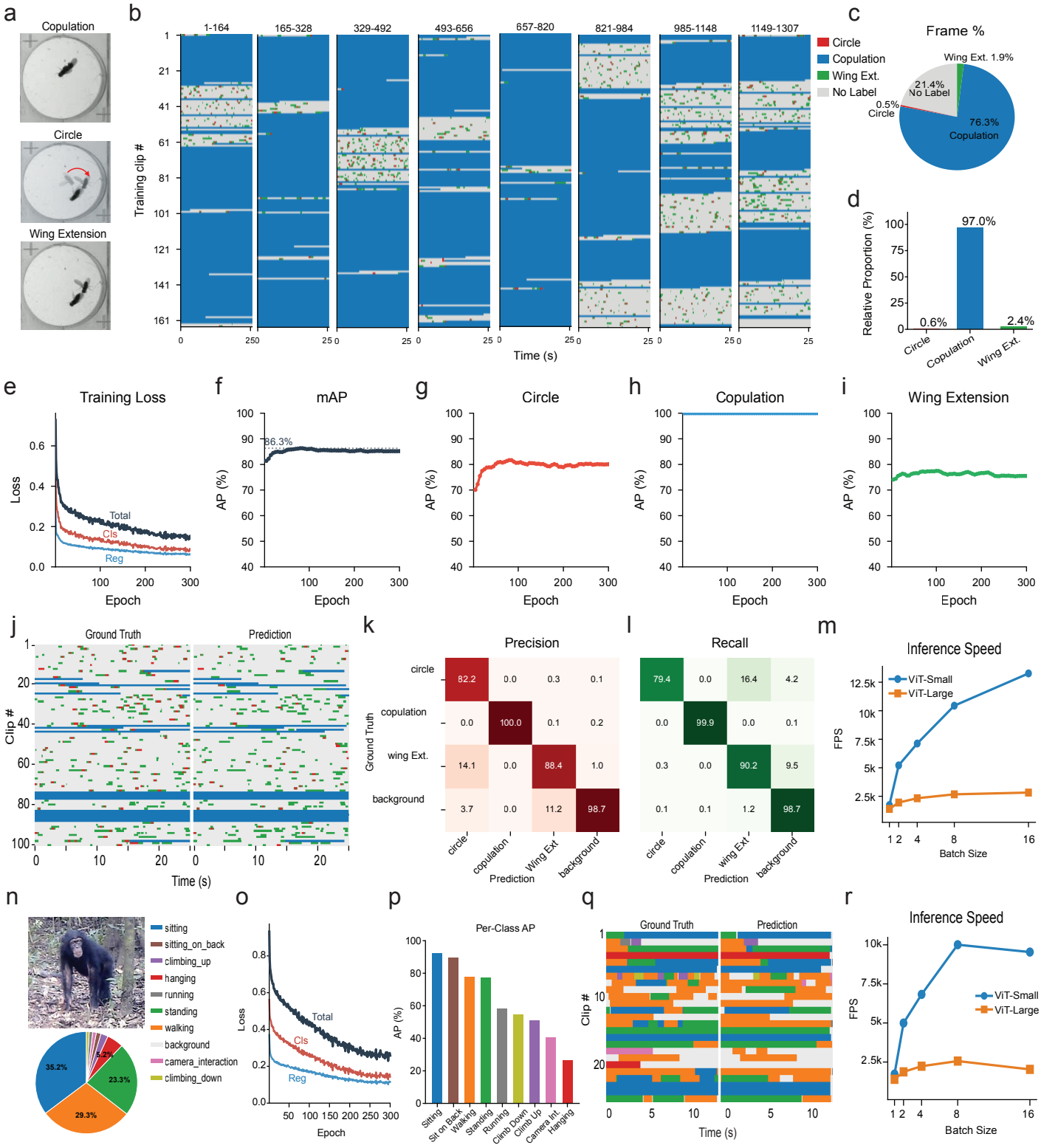

Figure 4