OPEN ACCESS

Neuroscience METHODS REVIEW

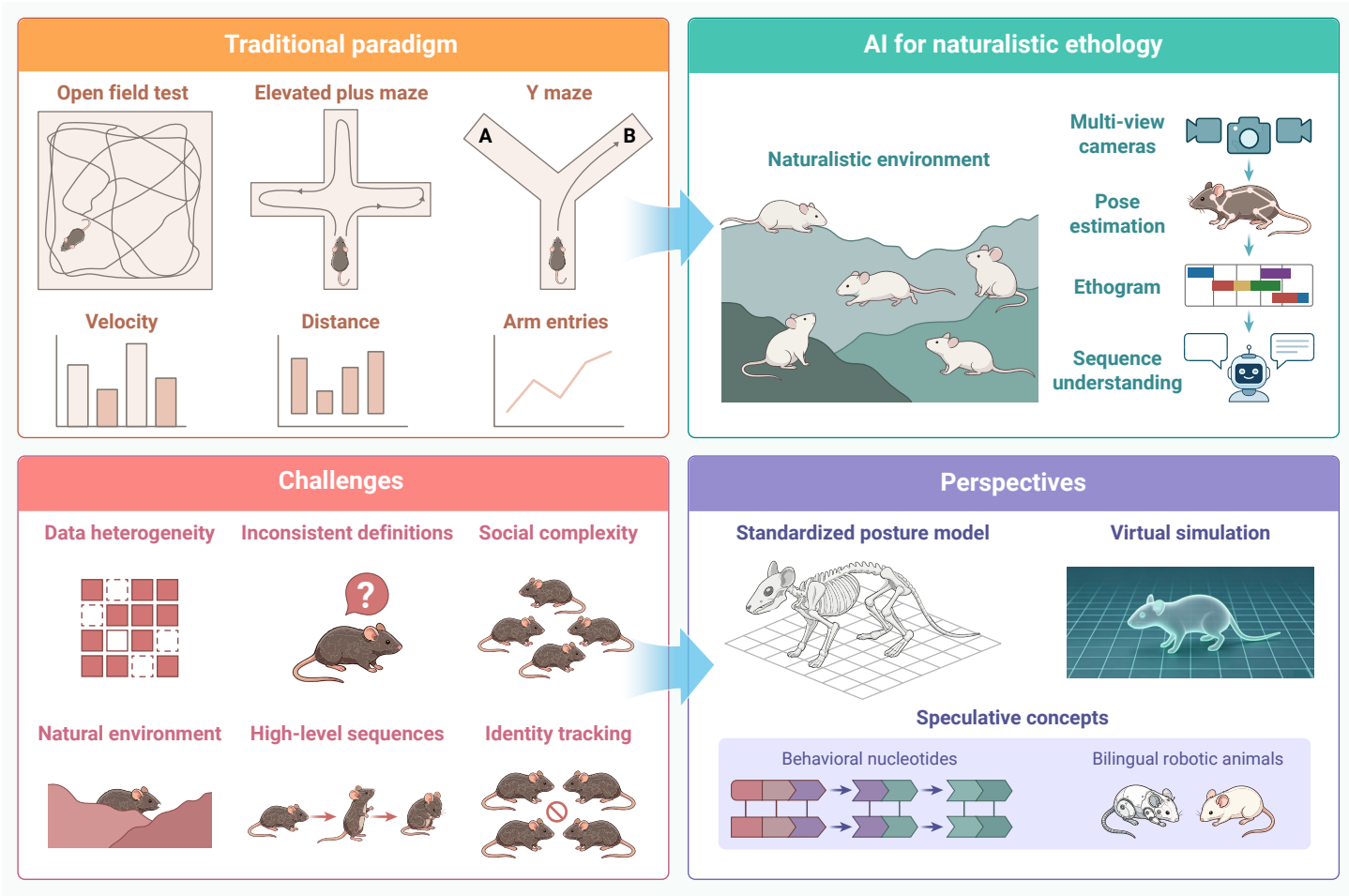
Technological advances in computational neuroethology

Ke Chen,[2,8] Yaning Han,[3,8] Kang Huang,[2,*] Liping Wang,[1,5,6,7,*] and Pengfei Wei[1,4,5,6,7,*]

*Correspondence: k.huang@bayonesci.com (K.H.); lp.wang@siat.ac.cn (L.W.); 101014012@seu.edu.cn (P.W.)

Received: April 14, 2026; Accepted: August 24, 2026; Published Online: September 3, 2026; https://doi.org/10.59717/j.tirv.2026.100012

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Traditional ethology often relies on task-constrained apparatuses, yielding low-dimensional behavioral measures.
- Artificial Intelligence (AI) now enables the analysis of freely moving animals in more natural spaces.
- AI drives pose estimation, action recognition, identity tracking, and body language understanding.
- Ambiguous definitions, data heterogeneity, and social complexity remain major challenges.
- Physical simulations and the forward-looking concept of "behavioral nucleotides" may inspire future research.

METHODS REVIEW Neuroscience

www.the-innovation.org/the-innovation-reviews

Technological advances in computational neuroethology

Ke Chen,[2,8] Yaning Han,[3,8] Kang Huang,[2,*] Liping Wang,[1,5,6,7,*] and Pengfei Wei[1,4,5,6,7,*]

[1]Shenzhen Key Laboratory of Neuropsychiatric Modulation, Shenzhen-Hong Kong Institute of Brain Science, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[2]Guangdong BayOne Biotech Co., Ltd., Guangzhou 510700, China

[3]Department of Pharmacology, Joint Laboratory of Guangdong-Hong Kong Universities for Vascular Homeostasis and Diseases, SUSTech Homeostatic Medicine Institute, School of Medicine, Southern University of Science and Technology, Shenzhen 518055, China

[4]School of Biological Science and Medical Engineering, State Key Laboratory of Digital Medicine, Southeast University, Nanjing 211189, China

[5]Guangdong Provincial Key Laboratory of Brain Connectome and Behavior, Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[6]SIAT-HKUST Joint Laboratory for Brain Science, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[7]State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[8]These authors contributed equally

*Correspondence: k.huang@bayonesci.com (K.H.); lp.wang@siat.ac.cn (L.W.); 101014012@seu.edu.cn (P.W.)

Received: April 14, 2026; Accepted: August 24, 2026; Published Online: September 3, 2026; https://doi.org/10.59717/j.tirv.2026.100012

Citation: Chen K., Han Y., Huang K., et al. (2026). Technological advances in computational neuroethology. *The Innovation Reviews* **1**:100012.

Animal behavior offers a critical window into brain function, yet its inherent spatiotemporal complexity has long posed a formidable challenge to objective and fine-grained quantification. This review systematically surveys how Artificial Intelligence (AI) is revolutionizing ethology. We first dissect the key advances in the AI-driven pipeline for behavioral quantification, from pixel-level pose estimation, to automated feature identification and classification, to identity tracking and social communication, and finally to semantic-level behavioral interpretation, achieving a new level of precision and objectivity compared to manual and traditional methods. Subsequently, we critically examine the persistent challenges, including the ambiguity of behavioral definitions, the lack of standardization across experimental paradigms, and the difficulty in understanding higher-order behavioral sequences. Finally, we discuss emerging directions such as virtual animals and speculative concepts, including "behavioral nucleotides". The advent of AI is not only refining experimental methodologies but also providing new conceptual and analytical tools for understanding animal behavior and strengthening its relevance to translational research.

INTRODUCTION

Animal behavior refers to the dynamic responses exhibited by organisms to adapt to their environment for survival and reproduction,[1,2] which can be regarded as an explicit phenotype of the brain.[3] These manifestations are regulated by internal physiological states, including gene expression and hormone levels, as well as by external environmental stimuli. For example, murine feeding behavior is internally modulated by hypothalamic AgRP neurons while being triggered externally by food cues;[4] similarly, instinctual defensive responses under the looming paradigm are shaped by specific neural pathways and constrained by spatial factors like the presence of a nest.[5,6] Behavior is a bridge between neural circuits and their environments,[7] making it essential for studying neural mechanisms and evaluating drug efficacy. Several frameworks in computational ethology describe behavior hierarchically, progressing from postural configurations and movement modules to higher-order behavioral sequences and comprehensive ethograms.[8-12]

The classical paradigm in ethology has long been predicated on manual observation and descriptive recording. In these highly controlled settings, most environmental variables are held constant to isolate specific behaviors, thereby reducing them to a few degrees of freedom.[13,14] This approach demands restrictive experimental designs, such as utilizing Y-mazes to calculate correct food-seeking attempts for memory evaluation, three-chamber tests to measure social preference, or open-field tests to count central zone entries for anxiety assessment. Consequently, the resulting analyses were often restricted to low-dimensional metrics such as trajectory and velocity. In recent years, research has gradually shifted toward behavior in more naturalistic and unrestrained settings, allowing complex brain states to manifest more fully and yielding more authentic behavioral expressions.[2] However, this pursuit of ecological validity dramatically escalates the spatiotemporal complexity of behavior. Manual observation in these unconstrained scenarios is highly subjective, labor-intensive, and prone to error, severely limiting high-throughput analysis. More importantly, traditional manual quantification no longer matches the high resolution and granularity of modern neural recording and manipulation technologies, such as high-density silicon probes and optogenetics. This discrepancy imposes far more stringent demands on the methods used for behavioral quantification.

Early attempts to automate tracking in the 2000s relied on traditional computer vision techniques like background modeling and edge detection. While useful for simple objects, these human-defined algorithms lacked the generalization needed for complex biological scenes. Over the past decade, deep learning[15] has catalyzed a technological revolution in computational ethology. The integration of techniques like millimeter-scale keypoint estimation, three-dimensional (3D) reconstruction, and robust identity tracking allows for the precise capture of each animal's complete kinematic profile.[16,17] On this foundation, spatiotemporal neural networks and behavioral low-dimensional embedding pipelines enable the objective and accurate classification of behavioral segments, including both interactive and individual behaviors, as well as the determination of their start and end times.[8,10,18-21] Unsupervised behavioral decomposition has begun revealing novel, sub-second movement subtypes previously invisible to human observers,[8] while supervised models provide highly accurate real-time classification.[22] Further, emerging Multimodal Large Language Models (MM-LLMs) may support zero-shot and interactive behavioral description and analysis.[23-25] Despite this rapid progress, significant hurdles continue to impede scientific advancement, including highly customized datasets, the semantic ambiguity of behavioral definitions, the difficulty of quantifying social interactions, challenges in capturing and interpreting behavior in naturalistic settings, the complexity of high-dimensional behavioral sequences, and ensuring unique animal identification. Looking forward, we discuss technology-supported directions such as standardized animal models and virtual environments, together with the forward-looking concepts of "behavioral nucleotides" and "bilingual robotic animals". This review systematically surveys the cutting-edge advancements in AI-driven behavioral analysis, critically dissects the current challenges, and outlines future directions and opportunities.

CURRENT ADVANCES

The paradigm shift toward naturalistic observations from limited behavioral tests

Traditional behavioral research has historically relied on highly structured apparatuses where experimenters fix most environmental parameters to retain minimal degrees of freedom. Researchers use tools like T-mazes, radial arm mazes, Y-mazes for evaluating spatial working memory, and three-chamber arenas for assessing social preference[13,14] (Figure 1A). These physical setups deliberately strip away environmental noise to isolate single variables. By forcing animals into a narrow set of predetermined choices, they succeed in standardizing the experiment but inevitably sacrifice the observa-

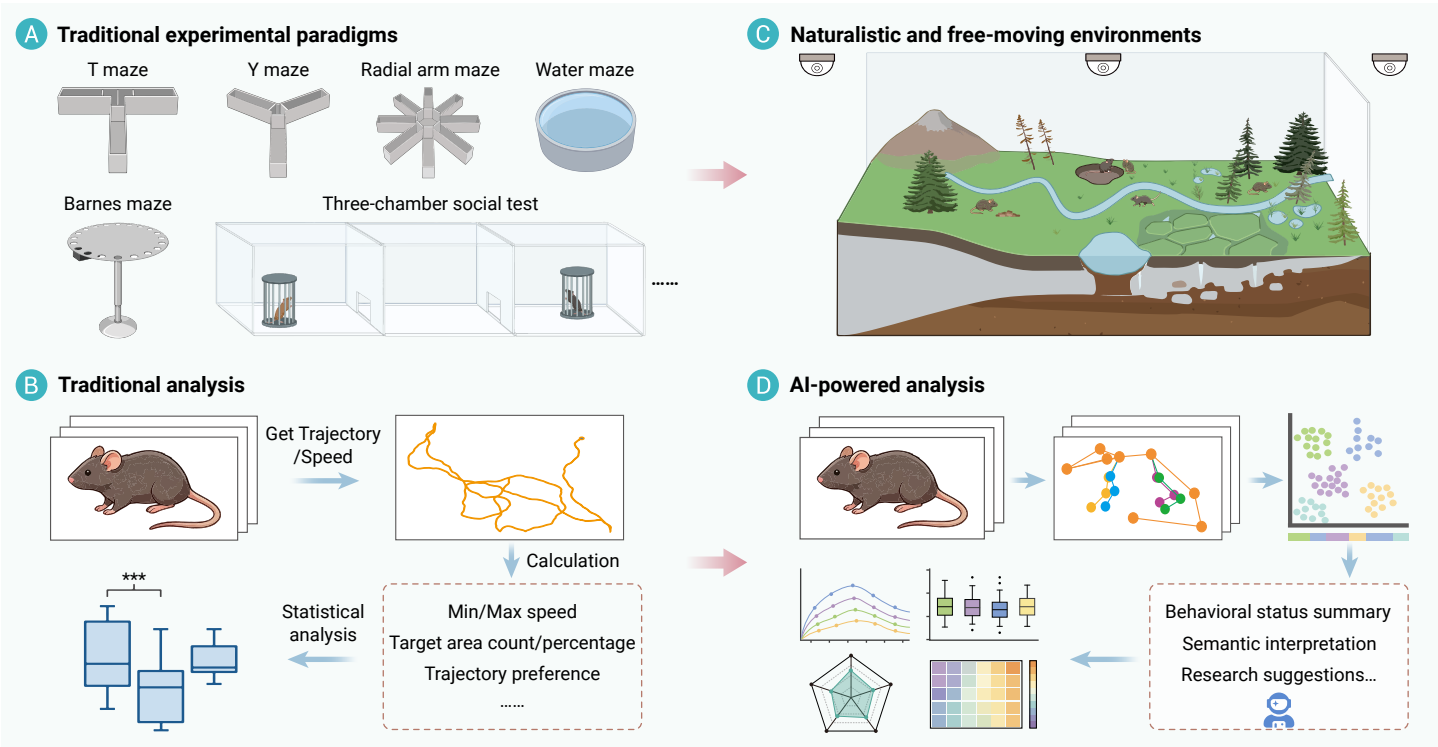


Figure 1. Paradigm shift from traditional behavioral assays to AI-powered naturalistic animal behavior analysis (A) Traditional experimental paradigms. Traditional research often relies on task-constrained apparatuses such as the T-maze, Y-maze, radial arm maze, water maze, Barnes maze, and three-chamber social test to measure specific, limited behavioral metrics. (B) Traditional behavioral analysis workflow, progressing from video capture to trajectory and speed extraction, calculation of basic behavioral metrics, and statistical analysis. (C) Naturalistic and free-moving environments allow animals to express richer behavioral repertoires, including feeding, social interaction, drinking, running, and hiding. (D) Five-stage AI-powered workflow comprising Capture, Extraction, Mapping, Understanding, and Analysis. Capture concerns the acquisition of raw visual and multimodal data; Extraction converts these inputs into structured representations such as poses, trajectories, identities, and three-dimensional reconstructions; Mapping identifies, clusters, or classifies behavioral patterns; Understanding integrates behavioral outputs with experimental context and domain knowledge to generate behavioral summaries, higher-level semantic interpretations, and research suggestions; and Analysis quantitatively examines the resulting representations and behaviors to derive statistically and biologically supported conclusions

tion of spontaneous, ecologically relevant actions. Confining the animal's physical repertoire in this way inherently limits the scope of quantifiable data and often fails to reflect the actual cognitive and physical capacity of the subject. In addition, conventional assays often require animals to be removed from their home cages and transferred by experimenters into unfamiliar apparatuses, potentially altering their behavioral state relative to the familiar home-cage environment and introducing handling-dependent variability into the measurements.[26]

Matching these physical restrictions, the analytical methods paired with these setups are equally constrained. Traditional workflows begin with video capture, followed by extraction of center-of-mass trajectories and speed. Researchers then derive rudimentary scalar values from these outputs, which serve as primary readouts for statistical testing (Figure 1B). Common examples include counting correct food-seeking attempts, measuring the percentage of time spent interacting with a novel conspecific, or recording the frequency of central zone entries in an open field to quantify anxiety levels. Although such simplified readouts facilitate high-throughput screening, they compress complex, multidimensional behavioral states into a limited set of low-dimensional metrics.

Treating a complex organism merely as a moving dot discards critical postural information. A center-of-mass trajectory cannot distinguish whether a stationary animal is freezing in fear, sleeping, or pausing to groom. It ignores spine curvature, limb coordination, and subtle head movements. Ultimately, these low-dimensional metrics fail to capture the nuanced motor dynamics necessary to link neural circuit activity with actual physical expression. More importantly, this gross simplification creates a severe mismatch with the granularity of modern neuroscience tools. We now have the technology to record from thousands of individual neurons simultaneously using high-density silicon probes, yet we often correlate this massive, high-resolution neural data with a single behavioral variable like speed or zone preference. Bridging the gap between brain and behavior therefore requires a fundamentally different observational framework.[7]

Driven by this necessity, modern ethology is undergoing a systemic shift toward unrestrained, naturalistic experimental arenas[2,3] (Figure 1C). In larger naturalistic arenas, animals can freely navigate complex terrains and express feeding, social interaction, drinking, running, and hiding behaviors. Such environments allow researchers to observe unconstrained behavioral flexibility that better reflects complex internal states. These open environments elicit continuous and diverse behavioral sequences, offering a more realistic window into how the brain processes and responds to real-world challenges. Alongside larger naturalistic arenas, automated home-cage monitoring provides a complementary route for continuous, long-term observation of spontaneous behavior while reducing repeated handling and experimenter intervention.[27,28]

Previous studies have proposed conceptual and computational frameworks for organizing animal behavior analysis. Building on these foundations, we organize recent advances in computational ethology and artificial intelligence into five operational stages: Capture, Extraction, Mapping, Understanding, and Analysis (Figure 1D). Capture concerns the acquisition of raw visual and multimodal data; Extraction converts these inputs into structured representations such as poses, trajectories, identities, and three-dimensional reconstructions; Mapping identifies, clusters, or classifies behavioral patterns; Understanding integrates behavioral outputs with experimental context and domain knowledge to generate behavioral summaries, higher-level semantic interpretations, and research suggestions; and Analysis quantitatively examines the resulting representations and behaviors to derive statistically and biologically supported conclusions. These stages are presented sequentially for organizational clarity, although they may interact iteratively in practical research workflows. The relationship between this organizational synthesis and representative prior frameworks is summarized in Table 1.

AI for animal pose estimation

To decode the rich behavioral repertoires elicited in naturalistic environments, modern ethology relies on a comprehensive AI-powered workflow.

Table 1. Relationship between the present five-stage organizational synthesis and representative prior frameworks in computational ethology

Framework	Primary emphasis	Correspondence to the present five-stage workflow	Scope relative to the present review
Anderson & Perona (2014)[11]	Establishing computational ethology as an objective and multiscale science of behavioral measurement, with links to neural activity and environmental context	Provides the broad conceptual foundation for Capture, Extraction, Mapping, and Analysis by emphasizing automated measurement, quantitative representation, and multiscale characterization of behavior	A foundational conceptual framework for computational ethology rather than a stage-wise organization of contemporary AI technologies
Wiltschko et al. (2015)[8]	Unsupervised decomposition of continuous three-dimensional behavior into recurring sub-second modules and characterization of their transition structure	Corresponds primarily to Extraction, Mapping, and Analysis through three-dimensional behavioral representation, unsupervised module discovery, and transition-probability analysis	A focused computational approach for discovering fine-scale behavioral modules rather than a complete workflow from multimodal acquisition to semantic understanding
Datta et al. (2019)[2]	Establishing computational neuroethology as an integrated field connecting naturalistic behavior, neural activity, causal manipulation, and computational modeling	Supports naturalistic Capture, quantitative behavioral representation, brain–behavior integration, and Analysis across multiple spatial and temporal scales	A field-level research agenda and conceptual foundation rather than an operational taxonomy of AI-based processing stages
Pereira et al. (2020)[9]	Quantifying animal motion and behavioral dynamics using modern tracking, pose-estimation, representation-learning, and machine-learning methods	Closely corresponds to Capture, Extraction, Mapping, and Analysis, particularly for pose tracking, behavioral representation, supervised classification, and unsupervised discovery	A methods-oriented synthesis centered on behavioral quantification; the present review extends this scope to multimodal acquisition, individual identification, social communication, and semantic Understanding with MM-LLMs and AI agents
Ours	Capture, Extraction, Mapping, Understanding, and Analysis	Organizes contemporary AI technologies according to their inputs, processing objectives, outputs, and roles within animal behavior research	Provides an updated operational synthesis spanning multimodal data acquisition, pose and three-dimensional reconstruction, behavioral mapping, identity and social analysis, semantic interpretation with MM-LLMs and AI agents, and quantitative downstream analysis

This pipeline begins with RGB scene capture in single- or multi-animal settings (Figure 2A). Optional multimodal sensors, including infrared cameras, depth sensors, and environmental monitors, can further enrich the acquired data (Figure 2B). Following this acquisition, the critical first computational step is extracting structured kinematic features. While traditional computer-vision tracking methods often lack robustness and generalizability in complex scenes, AI-driven pose estimation can extract structured kinematic representations while reducing sensitivity to irrelevant background variation.

Deep-learning-based animal pose-estimation toolboxes now provide transfer learning capabilities and high tracking accuracy across a range of experimental settings (Figure 3A). Among them, DeepLabCut[16] combines relatively limited manual annotation with user-friendly APIs, interfaces, and documentation, quickly establishing itself as one of the most widely used open-source pose estimation tools in the field. However, larger models demand greater computational resources and longer training and inference times, driving research into developing smaller, more efficient pose estimation systems.[29,30]

In multi-target scenarios (Figure 3A), pose estimation must address occlusion and keypoint association in addition to precise localization; two main strategies are widely used.[31] Bottom-up approaches first detect all keypoints in an image and then assemble them into individual targets. OpenPose[32] introduced Part Affinity Fields (PAFs) for accurate multi-target pose estimation and inspired subsequent methods.[31,33,34] Top-down approaches instead use multistage pipelines that reduce multi-animal tracking to single-animal pose estimation after initial detection or instance segmentation.[35,36] Creating multi-animal datasets remains time-consuming, especially for images containing three or more animals, although data augmentation and few-shot learning can reduce the annotation burden.[19,37]

Natural environments present additional challenges including rapid scene changes, intense motion, uneven or dynamic lighting. Appropriate solutions involve semi-supervised learning, few-shot learning, and domain adaptation.[38–40] Continuous dataset expansion also improves model performance in real-world applications.[41,42]

AI for animal 3D reconstruction

Although traditional two-dimensional (2D) images or overhead videos can support behavioral classification, they omit vertical spatial information and may lead to incorrect estimates of spatial relationships.[17] Three-dimensional (3D) reconstruction techniques (Figure 3A) can recover more complete animal morphology and provide richer inputs for motion measurement and behavioral analysis.[10] In ethology, 3D reconstruction can be grouped into three broad categories: keypoint reconstruction, dense point-cloud reconstruction, and 3D animal-shape modeling.

Keypoint reconstruction focuses on selected anatomical landmarks and offers high computational efficiency with modest resource requirements.[10] Dense point-cloud reconstruction captures the complete 3D surface and is particularly valuable for animals with highly deformable bodies.[43] Three-dimensional animal-shape modeling aims to create coherent representations of morphology. Once a parametric 3D mesh has been established, 3D shape can be inferred from 2D images using trained models or iterative optimization.[44]

Orthogonal to these representation types, 3D reconstruction methods can also be categorized by their computational pipelines into multistage and end-to-end approaches. Multistage methods first extract 2D keypoints or features from images or videos and then recover 3D coordinates using multiview geometry.[45] These methods are interpretable and require only 2D image annotations, but they depend strongly on multi-camera calibration and 2D tracking accuracy. End-to-end methods instead predict 3D animal skeletons directly with neural networks. These networks jointly learn keypoint extraction and 2D-to-3D mapping.[46,47] However, they can be sensitive to background variation and generally require larger datasets; the diversity of behavioral experiments and the burden of annotation may therefore limit their practical use.

AI for determining animal behavioral features, duration, and classification

In most studies aiming to understand behavioral patterns or conduct mechanistic or drug research, accurately determining behavioral features, temporal boundaries, duration, and categorical labels is essential for statistical analysis and mechanistic interpretation. Traditional manual scoring is limited by subjective bias, reaction-time variability, and inconsistent boundary definitions, particularly for sub-second actions or rapidly transitioning

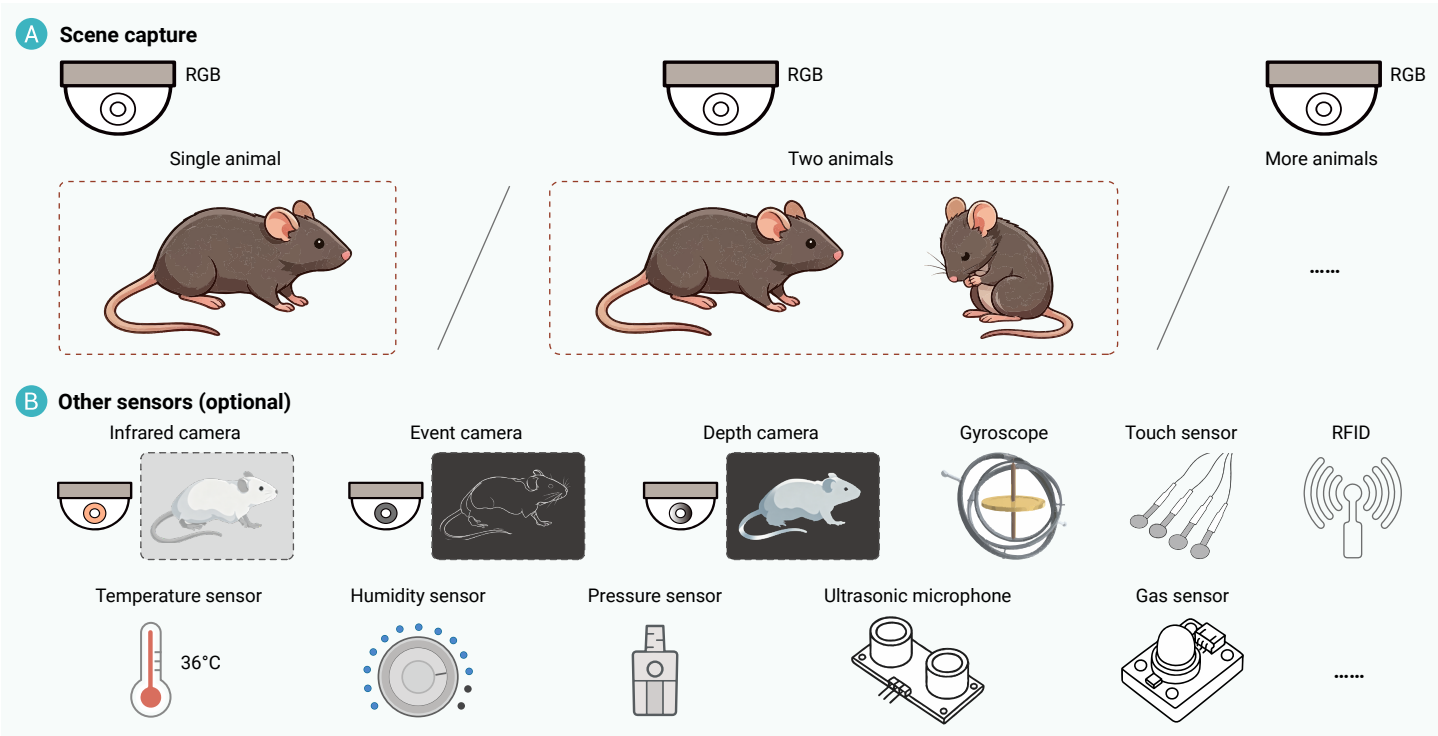


Figure 2. The capture stage in a typical AI pipeline for animal behavior analysis (A) Scene capture using RGB cameras tailored for single-animal, two-animal, or multi-animal environments. (B) Optional integration of other multimodal sensors to enrich the dataset, such as infrared cameras, event cameras, depth cameras, gyroscopes, touch sensors, RFID, and environmental sensors, including temperature, humidity, and pressure sensors, ultrasonic microphones, and gas sensors.

behaviors. AI approaches address these limitations by enabling automated extraction of high-dimensional kinematic features, objective determination of behavioral onset and offset frames, and scalable classification across large datasets. Broadly, AI-based behavioral analysis frameworks can be categorized into supervised and unsupervised learning paradigms, each offering distinct methodological strengths and application scenarios.[12]

Supervised learning (Figure 3B) uses labeled examples to predict predefined behavior categories. Conventional pipelines classify manually designed spatiotemporal features,[21] whereas deep-learning approaches can learn discriminative representations directly from video frames.[48] Although supervised methods support accurate and rapid prediction, including real-time classification,[22] they depend heavily on manually labeled datasets and may generalize poorly to unseen behaviors.

Unsupervised learning, a label-free approach, discovers latent patterns by analyzing the data's intrinsic structure, enabling automatic behavioral mapping (Figure 3B). Key steps in unsupervised behavior recognition include feature extraction and computation, similarity measurement, dimensionality reduction and embedding, and clustering.[9-10,18-19,49] These methods decompose continuous behavior into modular components, facilitating objective behavioral representation and enabling the discovery of previously unannotated behavioral motifs. Some unsupervised algorithms[10,19,50] dynamically adjust behavioral boundaries during runtime, supporting the precise determination of sub-second behavioral onset and offset times instead of relying on fixed sliding windows. This adaptive segmentation better reflects the natural temporal structure of behavior. Furthermore, employing techniques like hierarchical clustering allows the resulting behavioral components to align with the inherent hierarchical structure of behavior. Only limited human annotation may be required for downstream labeling and interpretation, substantially reducing the overall annotation burden.[51]

Although unsupervised methods can reveal recurrent behavioral motifs, computational separability alone does not establish their biological relevance. Robustness across sessions, individuals, and experimental conditions, together with explainable analysis of the features driving motif assignment, can help distinguish stable behavioral structure from recording or algorithmic artifacts. Biological interpretation can be further supported by convergent evidence from physiological states, environmental context, neural recordings, and, where feasible, causal perturbation experiments.[52-54]

From a practical perspective, many unsupervised discovery pipelines prioritize behavioral structure discovery rather than real-time prediction, and the resulting motifs may exhibit within-cluster heterogeneity or ambiguous boundaries. Supervised approaches may therefore be preferable when predefined labels and accurate online prediction are required. Researchers should select approaches based on specific requirements.

AI for animal identification and social communication

When analyzing group behavior in multi-animal scenarios, AI frameworks serve two closely related objectives: maintaining consistent individual identities (Figure 3A) and quantitatively characterizing social interactions. Reliable identity tracking ensures continuity of subject-specific behavioral trajectories over time, while social analysis focuses on modeling the dynamic relationships between individuals within a group. Together, these two components enable both longitudinal profiling of individual animals and structured quantification of inter-individual dynamics, including interaction sequences, spatial distance, and role differentiation.

Radio-Frequency Identification (RFID) is a widely adopted technique.[55] It is employed not only in research settings to reliably differentiate individuals,[22] particularly those lacking distinct visual cues, but also finds extensive application in livestock management for monitoring individual metrics such as feed and water intake.[55] However, the implantation procedure itself is invasive and the associated discomfort may potentially influence the animal's natural behavior. With continuous advancements in image processing algorithms, non-invasive identity tracking can be achieved by integrating multi-view feature analysis, inter-frame temporal continuity constraints, adaptive filtering, and matching algorithms.[56] When identity swaps occur, manual frame-by-frame correction is required. Deep learning has further improved tracking by learning animal-specific feature representations from large datasets.[19,36,57] However, appearance cues alone remain unreliable when visually similar animals overlap or temporarily disappear from view. Recent approaches increasingly incorporate representation learning, transfer learning, and identity-correction strategies to improve tracking robustness and identity preservation across challenging multi-animal recordings.[58-60] Collectively, these developments mark a shift from local appearance matching

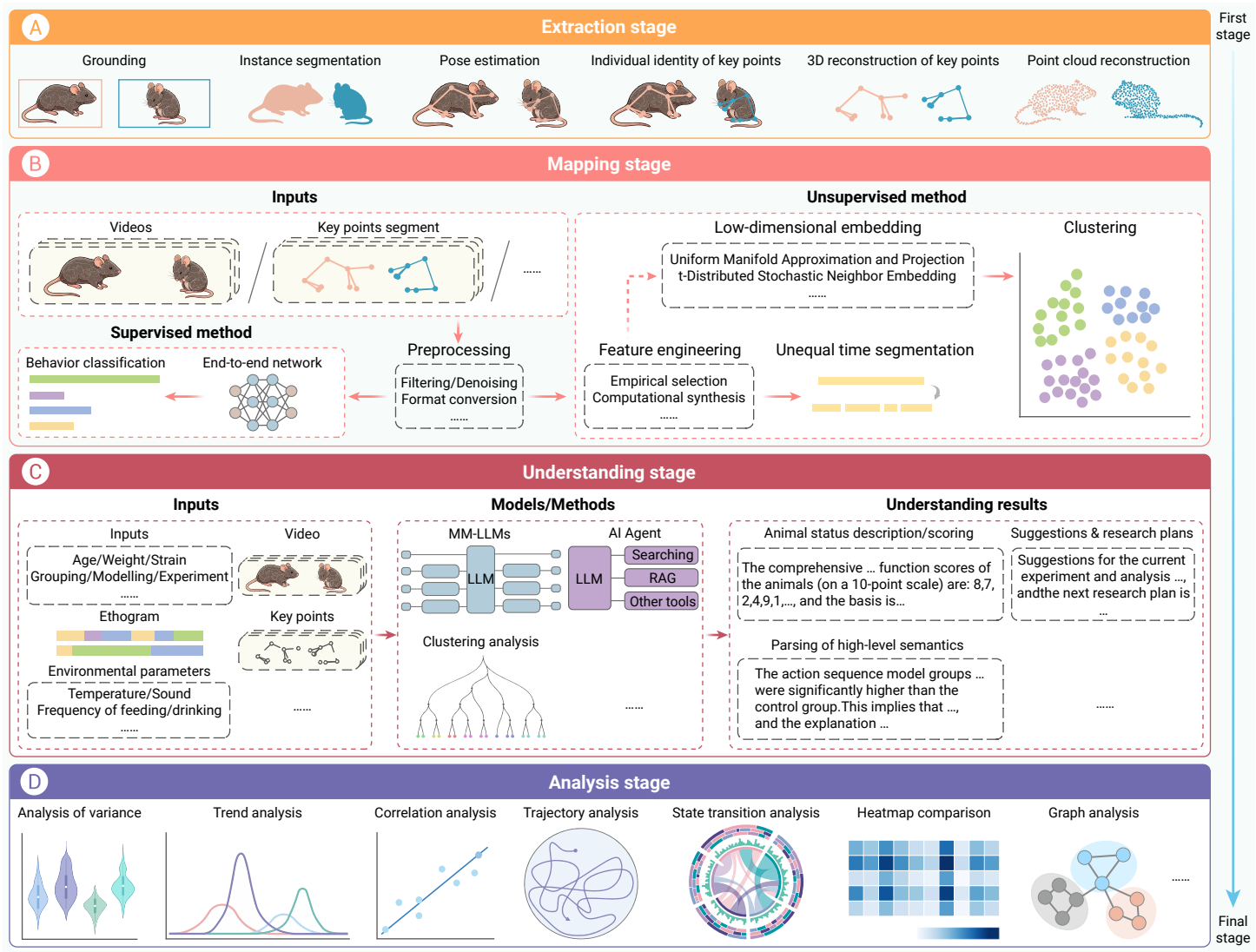


Figure 3. The processing and analysis stages in a typical AI pipeline for animal behavior analysis (A) The Extraction stage. This stage extracts simplified yet structured representations from raw inputs through tasks such as localizing animals (grounding), separating individuals (instance segmentation), estimating keypoint poses (pose estimation), maintaining individual identities over time (identity tracking), and reconstructing 3D pose or point clouds. (B) The Mapping stage. This stage classifies or discovers behavioral patterns using supervised end-to-end models or unsupervised methods such as low-dimensional embedding and clustering. (C) The Understanding stage. This stage employs advanced AI models and agentic systems, including Multimodal Large Language Models (MM-LLMs) and AI agents, to interpret behavioral outputs. By integrating contextual information and using strategies such as Retrieval-Augmented Generation (RAG), these systems may support higher-level semantic interpretation. (D) The Analysis stage. This final stage performs a wide range of quantitative analyses on the extracted features and classified behaviors, including analysis of variance, trend analysis, correlation analysis, and state transition analysis, to support statistically and biologically grounded conclusions.

toward more temporally integrated and identity-preserving tracking.

Multi-animal tracking approaches differ in their tracking targets, supervision requirements, computational demands, and validation settings. Table 2 summarizes representative multi-animal tracking systems, with reported performance interpreted in the context of each study's datasets, tasks, hardware, and evaluation metrics.

Robust identity tracking provides the necessary foundation for social quantification. In multi-animal environments, AI algorithms or systems can continuously measure the relative distances, orientations, individual behaviors, and interactive behaviors between identified individuals, converting them into computable features. Whereas supervised learning relies on predefined social behavior categories, unsupervised social quantification algorithms[19,47] extend pose-based modeling to multi-animal settings, enabling the generation of both individual ethograms and social ethograms that characterize interaction patterns between animals. AI can improve the accuracy, objectivity, and reproducibility of social quantification. On this basis, we can further investigate the dynamic rules of collective social interactions, as well as the underlying biological mechanisms including genes and neural circuits that regulate these social behaviors, and systematically characterize various autism-related social phenotypes in animal models.[47]

MM-LLMs and AI agents for animal behavior analysis and understanding

Transformer architectures established attention as a central mechanism for sequence modeling and laid the foundation for modern large language models.[23] Subsequent multimodal models extended Transformer-based architectures to align representations across text, images, video, and audio, enabling few-shot and zero-shot tasks such as question answering, comprehension, and generation.[61] Recent agent-based systems further extend large models toward multistep planning and interaction with external tools.[62]

Recent multimodal models have demonstrated strong capabilities in video understanding and broader multimodal reasoning tasks.[63,64] Ethological applications, however, require models to integrate observed behavior with contextual information such as species, strain, age, sex, health status, and experimental background.[65] Current general-purpose multimodal models increasingly integrate text, images, video, and audio,[66] yet ethology-specific behavioral features and their biological meanings remain underrepresented.

This limitation is particularly relevant because ethological interpretation depends on signals beyond visible movements and postures, including ultrasonic vocalizations,[67] tactile interactions, olfactory cues, and chemical signals. Physiological signals and environmental variables can further provide complementary information about internal states and behavioral context.

Table 2. Comparative summary of representative multi-animal tracking methods

Method	Scope	Annotation / input	Compute / throughput	Reported performance	Robustness / validation setting	Strengths	Limitations
multi-animal DeepLabCut (2022)[33]	2D multi-animal pose, identity tracking and re-ID	Body-part and instance labels; optional identity/appearance labels; video	Backbone/resolution dependent; lightweight tracking; no normalized end-to-end throughput	Ellipse MOTA 0.97 vs box 0.78; marmoset ID switches −26%; fish MOTA up to +10% with unsupervised re-ID	Mice, including parenting interactions; fish and marmosets; crowding/occlusion	Integrated pose assembly, tracklet stitching and re-ID; custom skeletons	Domain-specific labels; full occlusion or near-identical animals may need correction
idtracker.ai / new idtracker.ai (2019; 2026 update)[57,59]	Markerless appearance-based identity tracking	Raw video; no identity labels; 2026: self-supervised pairs, no all-visible segment	Video-specific learning; 2026 version up to 700× faster	Original often >99.9% IDF1 in easier videos; 2026 improved accuracy/occlusion robustness on 33 videos	Up to 100 animals; flies, zebrafish and mice	No tags; global identity learning; confidence/inspection tools	Needs reliable segmentation and stable appearance; difficult backgrounds, homogeneous animals and full overlap
SLEAP (2022)[31]	2D multi-animal pose and identity tracking	~200 labeled frames for ~90% peak accuracy; identity labels only for appearance-based ID	Up to 804 FPS; latency down to 3.2 ms	mAP 0.821 (flies), 0.774 (mice); ID switches: 0.91 (flies) and 22.7 (mice) per 100,000 frames	7 datasets; varied species, group sizes and environments	End-to-end workflow; top-down/bottom-up; real-time support	Temporal errors can propagate; appearance ID needs distinguishable animals; new domains need labels
AlphaTracker (2023)[35]	Top-down detection, pose and identity tracking; optional behavior clustering	Detection/keypoint labels; 600 train + 200 test frames; mAP >0.7 with 50 train frames	~2 h training on ~6,000 images; ~2 h for 7-min video (RTX 2080 Ti)	MOTA/MOTP/mAP: 82.2/86.2/87.2 for 2 mice; 84.0/87.2/85.6 for 4 mice	Home cages/chambers, angled views, webcams and head implants	Accessible animal-specific end-to-end pipeline; low-cost setups	Older YOLO/IoU/Kalman pipeline; slower than real time; close-contact errors and tuning
SBeA (2024)[19]	Few-shot multiview 3D pose, identity and social-behavior embedding	~400 multiview frames, single-animal pose labels and calibrated cameras; label-free identity after setup	Multistage 3D pipeline; no normalized throughput	Identity accuracy >90%; social-cluster purity >80%	Close-contact mice, parrots and dogs	Combines 3D tracking, identity and unsupervised social analysis	Requires calibrated multiview setup; multistage pipeline; dataset-specific metrics
ADPT (2025)[34]	Transformer anti-drift pose with multi-animal identity	Keypoint labels; mix-up social data; identity-recognition setup	Reported real-time capable; no standardized FPS	Identity accuracy: 93.16% for 10 unmarked mice; 90.36% for freely interacting mice, refined to 99.72%	Mouse, monkey, Drosophila and macaque; long/social recordings	Global context reduces drift; joint pose/identity estimation	Evidence stronger for anti-drift pose; limited severe-occlusion and long-term tracking tests
s-DANNCE (2025)[47]	High-resolution multiview 3D pose, identity/part assignment and social touch	Calibrated 6-camera video; semi-supervised 3D labels; anatomical constraints	Volumetric 3D CNN + GNN; high storage/compute; no standardized throughput	Close-interaction performance rivaled human labelers; >110 million 3D pose samples analyzed	Interacting rats/mice; close contact, occlusion and social touch	3D geometry, anatomy and GNN reduce chimeric part assignments	High acquisition/compute burden; mainly calibrated rodent settings
segTracker.ai (2024)[49]	Modular 2D detection, segmentation, pose, re-ID and tracklet association	Detection/segmentation/pose labels; 574 mouse, 36 ant and 33 fly images	Runs on modest-GPU laptop/PC; optical flow adds cost; no normalized throughput	Mouse inactivity IDF1 92.3; ant/fly >99.7% accuracy; mouse IDF1 up to 99.1	Mice, ants and flies; varied devices, video quality, interactions and inactivity	Model-agnostic modular pipeline; strong long-term tracking; web workflow	Depends on component models/training coverage; some comparisons use detector-derived ground truth
UDMT (2026)[60]	Annotation-free multi-animal tracking with transfer learning and bidirectional ID correction	Raw video + first-frame initialization; no train labels; unsupervised project fine-tuning	Per-project fine-tuning; no normalized throughput	Best or competitive HOTA, MOTA and IDF1 in the native comparisons	Mice, rats, Drosophila, C. elegans and Betta; crowding, occlusion, rapid motion and low contrast	No manual train labels; automatic refinement/tuning; bidirectional ID correction	Needs initialization/project adaptation; limited independent cross-lab validation

Table 3. Representative open benchmark datasets for AI-driven animal behavior quantification

Dataset	Primary task, species, and scale	Annotation protocol or ground truth	Evaluation protocol and metrics	Main limitation
AP-10K[41]	2D animal pose estimation; 10,015 images from 54 mammalian species in 23 families	Seventeen keypoints manually annotated and checked in COCO-style format; official training, validation, and test splits	OKS-based AP; supervised learning, human-to-animal transfer, and intra-/inter-family generalization	Still images without temporal identity or tracking annotations
APT-36K[71]	2D pose estimation and pose tracking; 2,400 clips × 15 frames from 30 species in 15 families	Seventeen keypoints, bounding boxes, and instance identities, with repeated manual checking and video-level data splits	OKS-based AP for single-frame pose estimation, unseen-species/family generalization, and pose tracking	Short 15-frame clips limit the evaluation of prolonged occlusion and long-term identity continuity
Rat 7M[46]	Single-animal 3D pose estimation; nearly 7 million synchronized RGB frames of freely behaving rats captured from up to six cameras	Marker-based motion-capture ground truth from 20 body sites synchronized with multiview RGB videos	Mean 3D landmark error and percentage of predictions below an 18-mm reference threshold	Rat-only laboratory setting and reliance on marker-based motion-capture ground truth
PAIR-R24M[42]	Multi-animal 3D pose estimation and social behavior analysis; 24.3 million RGB frames from 18 rat pairs and 24 viewpoints	Marker-based 3D ground truth for 12 landmarks per identified rat, together with 11 behavioral labels and 3 interaction categories	Mean per-joint position error or mean Euclidean error in millimetres, including behavior-stratified pose evaluation	Controlled rat-dyad setting with substantial storage and computational requirements
CalMS21[72]	Frame-level social behavior classification; mouse dyads with more than 1 million labeled and 6 million unlabeled pose frames	Seven tracked keypoints per mouse and expert frame-level behavior labels from a resident–intruder assay	Class-averaged F1 and mAP; standard classification, annotator-style transfer, and few-shot new-behavior tasks	Assay-specific behavioral taxonomy, class imbalance, and dependence on upstream pose tracking
AnimalTrack[73]	Multi-animal identity tracking in naturalistic videos; 58 sequences, ~25,000 frames, and 10 animal categories	Dense frame-level bounding boxes and individual identities manually annotated and inspected throughout each sequence	HOTA, MOTA, IDF1, ID switches, and related detection and identity-tracking metrics	Focuses on bounding-box-based identity tracking without pose or behavioral annotations

These heterogeneous data streams differ not only in their formats but also in their temporal and spatial properties. Video provides dense frame-based observations, vocalizations frequently occur as transient acoustic events, tactile interactions may be represented as sparse contact events, physiological signals are continuous time series with modality-specific sampling rates, and olfactory or chemical signals may vary more slowly and diffuse across space.

The key computational challenge is to align these heterogeneous signals with shared behavioral events or latent biological states. A multimodal framework may use modality-specific encoders to preserve the structure of each signal, timestamp synchronization and event-centered temporal windows to establish correspondence, and cross-attention, contrastive learning, or joint embedding strategies to learn shared representations across modalities. Because strict frame-level correspondence is often unavailable, asynchronous signals and missing modalities may require temporal resampling, interpolation, masking, uncertainty estimation, or models that do not assume exact temporal alignment. The resulting joint representation should preserve complementary modality-specific information while capturing shared behavioral and physiological states. MM-LLMs may subsequently integrate these aligned representations with contextual information, including species, strain, age, sex, health status, and experimental background, to support higher-level behavioral interpretation and report generation (Figure 3C). The effectiveness of such systems should be evaluated through cross-modal retrieval or prediction, robustness to missing or corrupted modalities, ablation of individual modalities, and the biological relevance of the learned joint representations.

Beyond multimodal interpretation, AI agents can extend large models from passive reasoning toward the coordination of multistep behavioral-analysis workflows. By decomposing research questions, maintaining contextual information, and invoking specialized models, analytical tools, or code, agents can link behavioral quantification with downstream analysis and interpretation.[62] This workflow-oriented capability is particularly relevant to ethology, where tracking, pose estimation, behavioral segmentation, statistical analysis, and interpretation are often interdependent (Figure 3A-D). Recent animal-behavior systems have begun to integrate these components within interactive or automated agentic workflows.[25,68-70]

Reliable ethological interpretation and agent-assisted behavioral analysis nevertheless remain emerging capabilities. Grounding model outputs and agent actions in observable behavior, domain knowledge, traceable evidence, and uncertainty estimates, together with expert review where appropriate, may improve the reliability of these systems.

Benchmark datasets and evaluation protocols

Standardized benchmarks provide a common basis for assessing whether advances in pose estimation, tracking, and behavioral classification generalize beyond a particular species or experimental setting. Available datasets span two-dimensional pose estimation across taxa,[41,71] multiview three-dimensional pose estimation,[42,46] frame-level behavioral classification,[72] and multi-animal identity tracking,[73] while differing in species coverage, recording conditions, annotation protocols, data partitions, and evaluation metrics. Table 3 summarizes representative open benchmarks across these task categories.

Despite these advances, direct comparisons across datasets remain difficult because anatomical keypoint definitions, behavioral taxonomies, train–test partitions, annotation procedures, and evaluation metrics are not yet standardized. More broadly, species-diverse benchmarks are most informative when their evaluation protocols characterize model performance and generalization across different species, individuals, and experimental conditions. Future benchmarks should therefore emphasize transparent annotation protocols, standardized data splits, annotation uncertainty, temporal identity continuity, and out-of-domain evaluation.

CHALLENGES

The dilemma between diverse experimental paradigms and uniform datasets

An experimental paradigm is defined as an integrated system encompassing operational procedures, technical tools, and quantitative metrics, systematically constructed to validate specific scientific hypotheses. In animal

behavior research, the selection of an experimental paradigm is typically dictated by the specific research objectives. To investigate the neural circuits underlying a particular disease, trace its progression, or explore the relationship between specific behaviors and physiological states, researchers often concentrate on predefined behavioral patterns or specific body parts.

Consider research using Parkinson's disease (PD) models[74] as an example. Investigators require varying depths of behavioral analysis tailored to their distinct research needs. For basic analyses, tracking a small number of forelimb keypoints may be sufficient to estimate pose and distinguish tremor from non-tremor movements. More detailed mechanistic studies may require denser keypoint definitions, higher spatial and temporal resolution, and finer behavioral categories, such as distinguishing normal grasping, tremulous grasping, and resting tremor. For studies focusing on bilateral motor coordination, implementing multiview systems might be required to resolve limb occlusion issues, necessitating the simultaneous tracking and analysis of keypoints on both forelimbs. Furthermore, when research extends to complex interactions between the forelimb and objects, the paradigm must integrate the spatial coordinates of target objects with existing keypoints and establish a dedicated pipeline for analyzing object interaction. While such goal-oriented research paradigms are precisely tailored to address specific scientific questions, they inherently lead to significant heterogeneity in dataset construction. Compounding this issue, research groups often develop customized experimental apparatus to address specialized scientific questions rather than employing standardized equipment.

This pervasive data heterogeneity creates substantial barriers to cross-study comparisons, rendering data from different laboratories often incompatible for direct integration or comparative analysis. Although standardization efforts are underway, they are constrained by practical challenges such as technical limitations, the time required, and inconsistent uptake among researchers. Community standards such as Neurodata Without Borders (NWB)[75] represent an important first step toward the standardized organization and integration of behavioral and neural data. By preserving synchronized timestamps, metadata, acquisition provenance, measurement units, and relationships between behavioral observations and neural recordings, NWB can improve data interoperability and facilitate cross-study analysis. Nevertheless, broader and more consistent adoption will require further efforts to overcome remaining technical limitations, reduce the time and effort required for implementation, and address uneven uptake across laboratories.

The issue of human consistency in behavior definition

A fundamental and critical challenge in contemporary behavioral research, particularly at the deep intersection with computational methods, is the lack of unified standards for defining behavior. This issue not only impedes the effective comparison and integration of findings across different studies but also poses significant barriers to the generalizability and reliability of automated behavioral analysis models.

First, regarding the delineation of specific behaviors, even seemingly simple and stereotyped actions, such as left versus right turns in mice, lack precise definition. For instance, over what duration and angular displacement does a turn qualify as a distinct left or right turn, as opposed to forward locomotion? Individual researchers may rely on subjective judgment based on their experience, leading to potentially biased experimental outcomes.

Second, the complexity of this problem is particularly pronounced for higher-level, more abstract behavioral concepts. Terms such as anxiety-like behavior, depression-like behavior, social exploration, and aggression may be defined differently across neuroscience, pharmacology, psychology, genetics, and specific experimental paradigms. Consider anxiety-like behavior: neuroscientists may quantify it using reduced time in or entries into the open arms of the Elevated Plus Maze, or reduced central activity and increased thigmotaxis in the open-field test. Although broadly labeled as anxiety-like behavior, measurements from different paradigms and species may capture distinct constructs, contributing to semantic ambiguity and uncertain scope. This definitional ambiguity is a major bottleneck when researchers integrate data across studies or develop computational models capable of recognizing such complex behavioral states.

Finally, it is essential to recognize the inherent limitations of relying solely on natural language descriptions to establish standardized behavioral definitions. Human language is often qualitative, subjective, and susceptible to the observer's experience, background knowledge, and contextual interpretation. Descriptions such as rapid movement, hesitant exploration, or slight twitch can be interpreted differently and assigned varying thresholds by different observers. Furthermore, behavior is intrinsically a dynamic process involving multidimensional spatiotemporal information such as velocity, acceleration, angle, and trajectory, which natural language struggles to convey with precision and efficiency.

For computer vision and machine learning algorithms that require precise quantification and pattern recognition, such ambiguity and subjectivity are particularly detrimental. These algorithms necessitate clear, objective, and quantifiable input features and annotations.

The difficulty in understanding social behavior

Social behavior in multi-animal environments is widespread and fundamentally important in nature, constituting the bedrock of animal social structures, group dynamics, information transmission, and collective decision-making. Consequently, the in-depth elucidation of the patterns, mechanisms, and functional significance of multi-animal social behavior holds indispensable value for fields including ethology, ecology, neuroscience, and evolutionary biology. However, as the number of interacting individuals increases, the complexity of social behavior increases combinatorially, presenting formidable challenges for behavioral computation and deep understanding. Currently, research on multi-animal social behavior indicates that larger group sizes tend to elicit a greater diversity and complexity of social interactions:

Dyadic Interactions (Two Animals): This is the most common and relatively well-studied context in traditional social behavior research. Behaviors such as chasing, sniffing, huddling, and mating between two mice, for instance, typically exhibit relatively distinct motor patterns. Nevertheless, defining social events can remain ambiguous; for example, what constitutes a valid approach behavior? From what distance does it commence?

Triadic Interactions (Three Animals): When the interactive system involves three individuals, the complexity of social dynamics increases significantly, far exceeding the simple aggregation of three independent dyadic interactions. Novel behavioral patterns, potentially exclusive to triadic contexts, may emerge. Examples include coalition formation, where two individuals align against a third; mediation, where one individual intervenes in a conflict between two others; audience effects, where an individual's behavior is altered by the presence of a third party; and competitive interactions, referring to the dynamic interplay among three individuals vying for resources or mating opportunities. How can these triadic-specific interaction patterns be clearly defined? How can they be distinguished from independent dyadic interactions? These questions currently lack standardized operational definitions and widely accepted analytical frameworks. This poses heightened demands on computational models, necessitating not only the simultaneous tracking and identification of the identity and pose of all three individuals but also the capacity to understand their more complex spatiotemporal relationships and conditional dependencies to infer these higher-order interaction patterns.

Group Interactions (N > 3 Animals): In groups comprising more individuals, the number of potential interaction combinations increases dramatically. Coupled with triadic, tetradic, and higher-order associations, this renders the social network structure exceptionally complex. At this scale, the research focus often expands from specific, direct inter-individual interactions to the overall properties and dynamics of the group. These include collective movement patterns such as swarming, the formation and dissolution of subgroups, and the stability and fluctuations of social hierarchy structures. Defining these group-level behaviors or states becomes considerably more challenging. For instance, how should a subgroup be delineated? Is mere spatial proximity sufficient, or are specific interaction frequencies or behavioral patterns requisite? How can collective motion be precisely quantified? For behavioral quantification models, this necessitates not only more powerful multi-target tracking capabilities but also advanced feature extraction and model construction techniques capable of effectively capturing and representing collective behavior.

Analogously, human social interactions demonstrate a similarly rapid increase in complexity with the number of participants. Transitioning from dyadic conversations to multi-party meetings and further to large-scale crowd gatherings, qualitative shifts occur in communication patterns, decision-making processes, emotional expression, and social dynamics. The diversity and species-specific nature of animal social behavior, together with limitations in research methodologies, including the inability to directly access linguistic information, make the understanding and quantification of animal social behavior challenging.

Capturing and understanding animal behavior in complex scenarios

While AI-driven behavioral analysis has achieved considerable advancements in relatively controlled laboratory settings, its application in complex, semi-natural, or entirely naturalistic environments faces a series of formidable challenges.

In contrast to the uniform illumination, clear backgrounds, and predictable spatial configurations typical of standard laboratory enclosures, complex environments are often characterized by variable lighting conditions. Natural light cycles such as diurnal variations, shadows cast by objects or other animals, and the intricate reflections and refractions from water bodies can drastically alter image quality and consistency. This, in turn, confounds algorithms trained on well-illuminated, stable image data, rendering robust feature extraction for pose estimation and identity recognition particularly difficult.[76]

Beyond these visual and environmental variations, the reliable capture of animal behavior data is further hampered by challenges related to both subject visibility and data acquisition systems. On one hand, animal subjects are frequently partially or completely occluded by vegetation, environmental enrichments, burrows, or uneven terrain. Such occlusions not only impede continuous tracking and accurate pose estimation but also make reliable animal segmentation from the background, a crucial prerequisite for many automated analysis pipelines, exceedingly challenging. On the other hand, while larger and more complex scenes may necessitate sophisticated multi-camera systems for adequate coverage, the deployment, calibration, and synchronization of these systems, as well as the fusion of multiview data, are considerably more intricate in such unstructured environments compared to constrained laboratory setups. These factors collectively contribute to the increased difficulty in obtaining high-quality, continuous behavioral data.

Compounding these challenges, the inherent complexity and unpredictability of the behavioral repertoire itself are amplified in complex scenarios. Animals may exhibit a broader spectrum of behaviors,[7] including camouflage, predation, competition, climbing, nesting, and cooperative group behaviors. These behaviors may manifest similarly to or entirely differently from those observed in standard laboratories and often display high spatiotemporal variability and multiscale interaction dynamics. Moreover, interspecific variations in responses to identical stimuli are common, and many naturalistic behaviors are notoriously difficult to annotate at scale, rendering their quantification exceedingly challenging.

In essence, while the objective is to study animals in environments more akin to their natural habitats to elicit a broader range of ecologically relevant behaviors, it is precisely this environmental complexity, behavioral diversity, and inherent unpredictability that pose substantial technical impediments to current analytical methodologies.

The difficulty in understanding high-level behavioral sequences

Rodents use body-language cues, including gestures and postures, to guide social communication and responses.[77] This sensitivity to social signals also extends to interactions with rat-like robots: rats have been reported to respond to robot help-seeking signals and to alter their emotional states in response to learned robot interaction patterns.[78,79] Research into rodent body language, however, remains in its nascent stages. Currently, knowledge is largely confined to certain overt stereotyped behaviors, such as the darting of female rats during courtship or fighting to display aggression and establish dominance. However, interpretation becomes considerably more difficult at higher levels of organization, particularly for complex behavioral sequences or strategies formed by concatenating multiple behavioral segments[80] according to specific temporal logic. This difficulty primarily stems from several factors:

A fundamental challenge is that animals cannot directly report the intent, motivation, or adaptive significance of higher-order behavioral sequences. During breeding, for example, a bird may forage, build a nest, sing, repel competitors, incubate eggs, and rear chicks. The function of each component may be apparent, but the rules governing their prioritization, necessity, and context-dependent adjustment remain difficult to infer. Researchers therefore reconstruct the organization of these sequences indirectly from ecological context, outcomes such as reproductive success or offspring survival, and associated physiological data. Such inference remains incomplete and may be influenced by observer expectations or anthropomorphic interpretation.[51,81]

Closely related to this interpretative difficulty is the complexity of the underlying neural mechanisms. Complex higher-order behaviors imply hierarchical organization, coordinated activity across multiple brain regions, and pronounced spatiotemporal dynamics.[2,82] Researchers must not only simultaneously monitor and analyze activity patterns across multiple brain regions at various temporal scales but also uncover how these activities integrate information, lead to decision-making, and ultimately drive and regulate the fluid execution of behavioral sequences. Furthermore, the information flow between these brain regions is bidirectional or even multidirectional, and their activity patterns change dynamically on millisecond-to-second timescales relevant to behavior. Consequently, decoding these patterns and precisely linking them to the generation, selection, and adjustment of specific behavioral strategies is exceedingly complex.

Another major obstacle arises from the difficulty of systematically decoding complex, multimodal communication signals. While substantial progress has been made in understanding specific communication signals, including alarm calls in primates,[83] the systematic deciphering of complex, multimodal body language in higher animals remains a formidable challenge. Higher vertebrates, with their greater bodily degrees of freedom, can generate an extensive repertoire of subtle and combined movements, postures, and gestures. This combinatorial complexity means that the vocabulary and grammar of their body language are far more extensive and nuanced than simpler signals, making systematic analysis and interpretation exceedingly difficult. Moreover, these complex bodily displays are often produced in conjunction with vocalizations or other sensory cues, yet research has rarely integrated these multimodal signals to understand their combined communicative intent or synergistic effects, leaving a significant gap in our comprehension of their social communication.

In summary, comprehending higher-order animal behavior is hampered by three primary challenges: the opacity of intent behind complex behavioral sequences due to the lack of direct animal reporting; the intricate and dynamic neural underpinnings that orchestrate these sequences; and the significant difficulty in systematically deciphering the rich, combinatorial, and often multimodal nature of sophisticated animal communication, particularly the complex body language of higher vertebrates and its interplay with other signaling modalities.

The problem of uniquely identifying individual animals

Achieving continuous and accurate individual identification and tracking of multiple cohabiting animals in complex scenarios remains a formidable challenge, even with advanced deep learning methods. The accuracy levels attained frequently prove insufficient for rigorous scientific inquiry, particularly in long-term, continuous monitoring tasks.

One major source of difficulty arises from the limited visual distinguishability between individuals. Humans are relatively easy to differentiate not only because of external variations such as clothing, hairstyles, and facial features, but more importantly due to stable skeletal structures and distinctive biometric identifiers like fingerprints and irises. In contrast, many commonly studied animals, including rodents such as mice and rats or social insects such as ants, exhibit minimal inter-individual variation in appearance. The use of inbred strains or littermates in experimental designs, intended to control for variables,[84] further exacerbates this visual homogeneity, making it exceedingly difficult for vision-based computational algorithms to identify stable distinguishing features. In rodents, the challenge is compounded by the lack of prominent skeletal landmarks, as well as their highly flexible bodies and

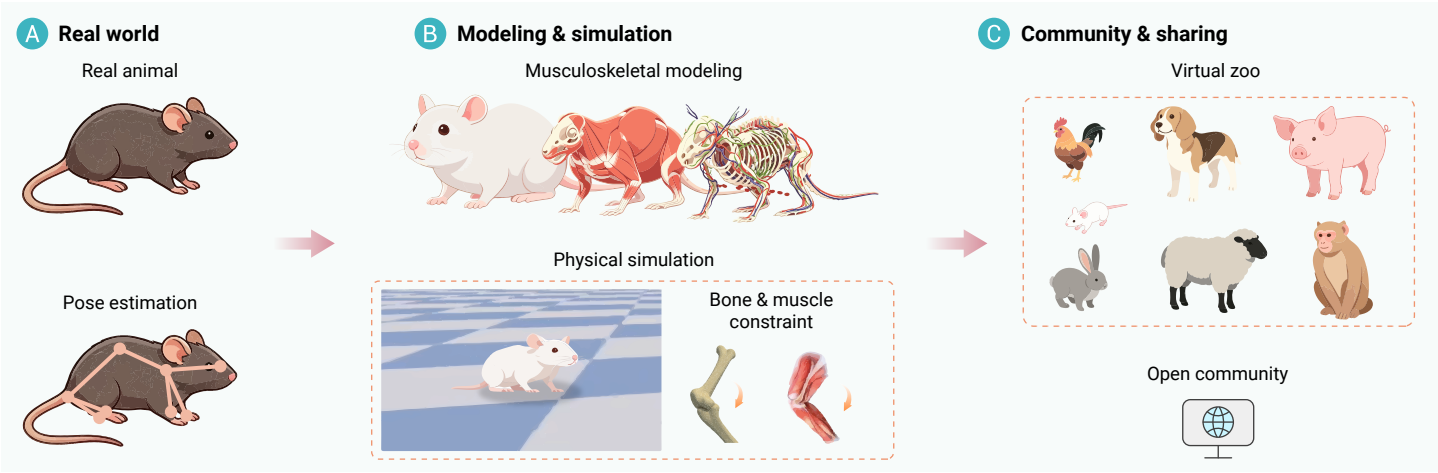


Figure 4. The development of a multispecies standard posture physical model (A) Real-world data acquisition, extracting pose estimates from real animals. (B) Modeling and simulation, where a standardized musculoskeletal model is used in a physics engine constrained by bone and muscle properties to generate physically plausible movements. (C) Community and sharing, where these models can be shared in an open community to create a virtual zoo, reducing redundant work and accelerating research.

spines, which allow a wide range of postures. Subcutaneous fat and muscle can further obscure subtle morphological differences.

Natural behaviors can also alter appearance over time. Grooming and curling change body contours; gnawing may cause localized fur loss or skin damage; and drinking or moving through water may change fur texture, color, and reflectance. Physiological color changes may occur in some species. These variations reduce the stability of appearance cues and increase ambiguity for identification algorithms.

Multiple animals sharing an environment frequently interact, producing severe body occlusions in which algorithms may capture only partial views or lose individuals entirely.[85] Identification under these conditions is highly error-prone. When paths intersect or animals separate after prolonged close contact, systems that lack stable identity features are particularly susceptible to identity swaps. If an uncorrected swap persists, all subsequent subject-specific analyses may be assigned to the wrong individual and become unreliable.

The high degree of visual similarity among animal individuals, the dynamic variability in appearance induced by their natural behaviors, and the pervasive severe occlusions inherent in dense multi-target interaction scenarios collectively constitute the core bottlenecks confronting current vision-based techniques for multi-animal individual identification and continuous tracking.

Recent methods can reduce identity loss under occlusion,[58–60,86] but prolonged full-body occlusion remains fundamentally ambiguous, particularly when visually similar animals engage in close and abrupt interactions. In such cases, multiple identity assignments may remain compatible with both the observed appearance and plausible motion trajectories. Recent methods therefore reduce, rather than eliminate, the risk of identity loss, and long-duration animal recordings may still require uncertainty reporting and targeted manual verification.

PERSPECTIVES

Multispecies standard posture physical model

A key future direction in ethological research is the establishment of more precise, standardized animal posture representation systems endowed with physically grounded properties (Figure 4). However, given the substantial interspecies differences in physiological structures, constructing a universal posture framework applicable to all species is exceptionally challenging. Instead, standardization can be pursued on a species-by-species basis.

Implementing this species-specific approach initially involves establishing species-specific standardized posture frameworks while maintaining format consistency across different species. Real-world animal data can first be used to extract species-specific pose estimates (Figure 4A). The standardized framework should then be grounded in anatomical and physiological information, including the number of bones, joint types, muscle attachment sites, and characteristic surface landmarks. For example, CT-derived anatomical models have been used to establish standardized skeletal and joint representations for mice.[87] Similar species-specific procedures could be extended to other animals when suitable anatomical data are available. Building upon this, output formats should be designed for uniform readability, such as consistently using JSON and adhering to a common set of definitions. This facilitates subsequent data integration and comparative statistical analysis across different species and individuals.

Once these standardized postures are established, the next critical step involves integrating biomechanical properties and physical constraints imposed by the environment.[87,88] For example, posture estimates can be refined by incorporating species-specific biomechanical parameters, including joint ranges of motion such as maximum flexion, extension, and rotation; muscle-force models defining maximum force and contraction velocity; and body-segment properties such as mass distribution, volume, and collision characteristics. These refined models can then be simulated or driven within established physics engines (Figure 4B). Advanced open-source pipelines have been developed to train neuromechanical control models directly from animal motion data.[89] These computational tools enable the precise simulation of experimental setups and facilitate deep neuromechanical behavioral analysis. By bridging the gap between kinematic tracking and physics-based simulation, such approaches illustrate the feasibility of grounding postural estimates in realistic biomechanical constraints. The resulting posture estimates could therefore be constrained by physical laws, helping to exclude overtly implausible estimates such as body-part interpenetration, self-intersection, or anatomically disordered limb configurations. Notably, while this approach more closely emulates the complex processes underlying physically grounded posture interpretation, it would inherently demand greater computational resources.

To support the computational demands of these simulations and accelerate field-wide progress, fostering an open community to build a multi-species physical posture model repository is crucial (Figure 4C). Researchers should be encouraged to upload, download, and iteratively refine pre-trained standard posture models, complemented by comprehensive user tutorials. This approach would mitigate redundant model development efforts, allow for continuous updates and optimizations, and thereby significantly accelerate research progress in the domain.

Through these three concerted efforts, a multi-species standardized physical posture framework can be progressively established, laying a robust foundation for quantitative, fine-grained, and intelligent research in animal ethology.

Virtual scenario training for AI real-world application

Currently, a large amount of data used for training AI models in animal behavior analysis is predominantly derived from real animal experiments. However, the acquisition of real-world data confronts numerous challenges: research involving real animals is resource-intensive, while recent NIH initiatives increasingly emphasize human-based research technologies and the

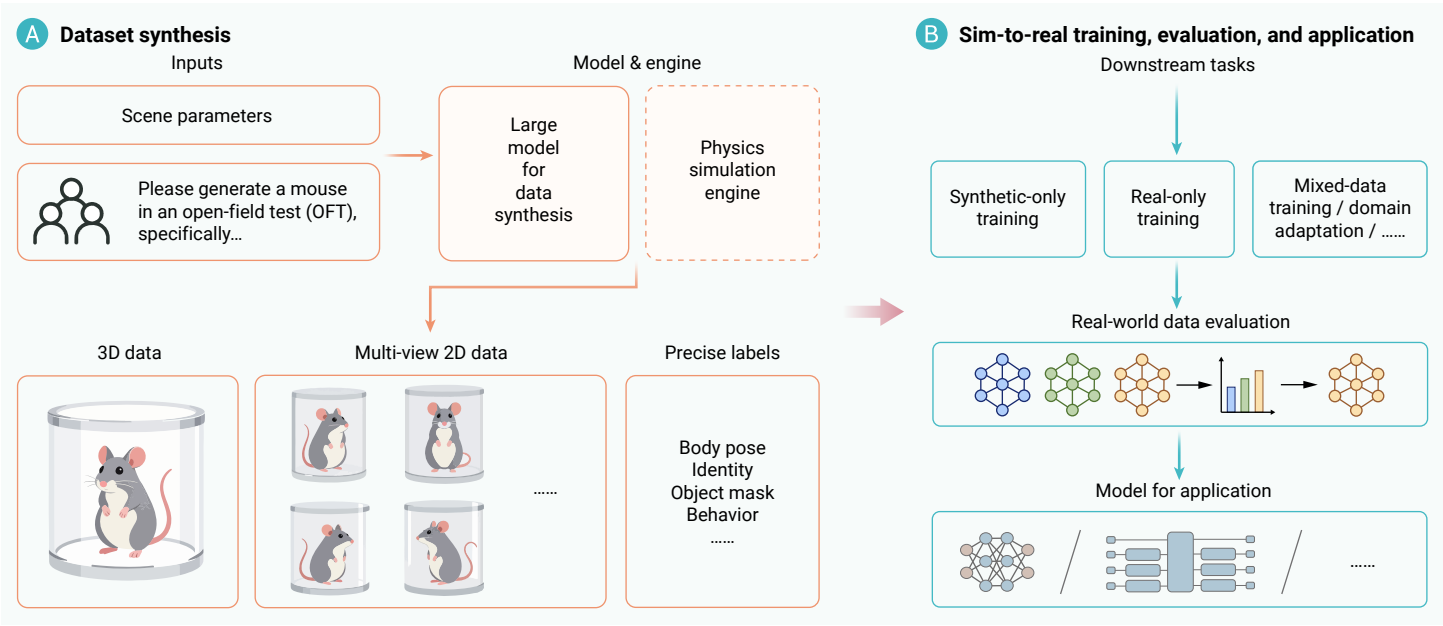


Figure 5. Simulation-based data generation and sim-to-real training, evaluation, and application for animal behavior analysis (A) Dataset synthesis. Scene parameters and natural-language task specifications are provided to a large model for data synthesis, which may interpret the requested experimental setting and invoke a physics simulation engine to generate three-dimensional data, multiview two-dimensional data, and precise, automatically aligned labels for body pose, identity, object masks, and behavior. (B) Sim-to-real training, evaluation, and application. For downstream tasks, synthetic-only, real-only, and mixed-data or domain-adaptation training strategies are evaluated using independent real-world data. A selected model can then be used for the target application.

reduction of animal use;[90] the duration of behavioral experiments is often protracted; and collecting high-quality data in complex naturalistic settings presents practical difficulties. Consequently, the ability to construct behaviorally realistic virtual animals and interactive virtual environments to generate training data (Figure 5) could alleviate some of these constraints and may support efforts to reduce animal use where scientifically appropriate.

The core of this goal is to create high-fidelity virtual worlds that generate synthetic data for AI training, an approach already used for algorithm validation and iterative optimization in autonomous driving[91] and robotics.[92] This strategy offers several advantages. It can generate large training datasets, helping to address data scarcity in real experiments and support data-intensive deep-learning models. Virtual environments can also simulate diverse environmental conditions and interaction patterns, which may improve model adaptation and generalization. In addition, synthetic data can provide automatically aligned ground-truth labels,[87] reducing manual annotation and its associated variability. At a more integrated system level, a large model for data synthesis could interpret natural-language task specifications together with structured scene parameters, invoke rendering or physics-simulation engines, and generate multimodal synthetic data with precise, automatically aligned labels (Figure 5A).

Current simulation-based animal research broadly follows two complementary routes. One focuses on synthetic visual data, using anatomically grounded three-dimensional animal models, constrained animation, rendering, and image-domain translation to generate realistic videos with automatically aligned pose labels.[87] The other uses physics-based neuromechanical simulation to reproduce interactions among neural control, biomechanics, and the environment. Platforms such as MuJoCo and GPU-parallelized MuJoCo-MJX/JAX, together with detailed musculoskeletal models, can support motion imitation, pose-to-model registration, and the learning of control policies from recorded animal movements.[88,89,93]

Sim-to-real design should be tailored to the downstream task (Figure 5B). For image-based pose estimation or behavioral classification, domain randomization may vary camera geometry, viewpoint, illumination, background, texture, occlusion, motion blur, and sensor noise. Neuromechanical applications may instead vary body dimensions, inertial properties, joint constraints, muscle or actuator parameters, contact friction, and initial conditions. Synthetic pretraining can subsequently be combined with limited real data through fine-tuning, image-domain translation, or unsupervised domain adaptation.[40,87] Because existing studies evaluate different objectives, their reported quantitative results are not directly comparable. Synthetic visual training has achieved pose-estimation performance comparable to typical manually annotated training data.[87] Neuromechanical pipelines have reported median errors of approximately 5 mm for pose-to-model registration and 10 mm for motion reproduction.[89] Shared real-world test sets comparing synthetic-only, mixed-data, and real-only training would provide a clearer basis for future evaluation.

Behavioral nucleotides and behavioral sequence forecasting

A longstanding idea in complex-systems research is that complex systems often exhibit modular and hierarchical organization.[94] Applying this idea to ethology,[8] we propose the forward-looking concept of "behavioral nucleotides" (Figure 6). Animal behavior emerges from continuous interactions among neural activity, biomechanics, sensory feedback, internal states, and the environment, whereas computational analysis generally requires these dynamics to be sampled and represented numerically. Analogous to how nucleotides in deoxyribonucleic acid (DNA) form sequences that encode genetic information, behavioral nucleotides are envisioned as elementary computational units for representing behavior, with their possible forms constrained by animal morphology, biomechanics, and environmental context (Figure 6A). Each sampled time point, represented by a multidimensional feature vector, can be treated as a candidate behavioral nucleotide (Figure 6B). Ordered combinations of these nucleotides over different temporal spans can form behavioral sequences and behavioral patterns.

Behavioral nucleotides may provide a new computational perspective for addressing several pervasive issues in contemporary animal behavior research, including ambiguous behavioral boundaries, subjective definitions, and a lack of standardization. If reproducible candidate units can be identified, behavioral trajectories may be represented as sequences of these units and analyzed using computational methods analogous to those used in bioinformatics, including sequence alignment, motif discovery, and clustering.[95,96] Related computational sequence models[97] may facilitate large-scale processing and pattern discovery, but their usefulness will depend on whether candidate units can be defined reproducibly. If stable candidate units and transition patterns can be identified, the same sequential structure may also support probabilistic, context-dependent short-horizon forecasting rather than deterministic prediction (Figure 6C). At the same time, representing continuous behavior as sequences of sampled computational units may simplify some aspects of the underlying dynamics, particularly gradual tran-

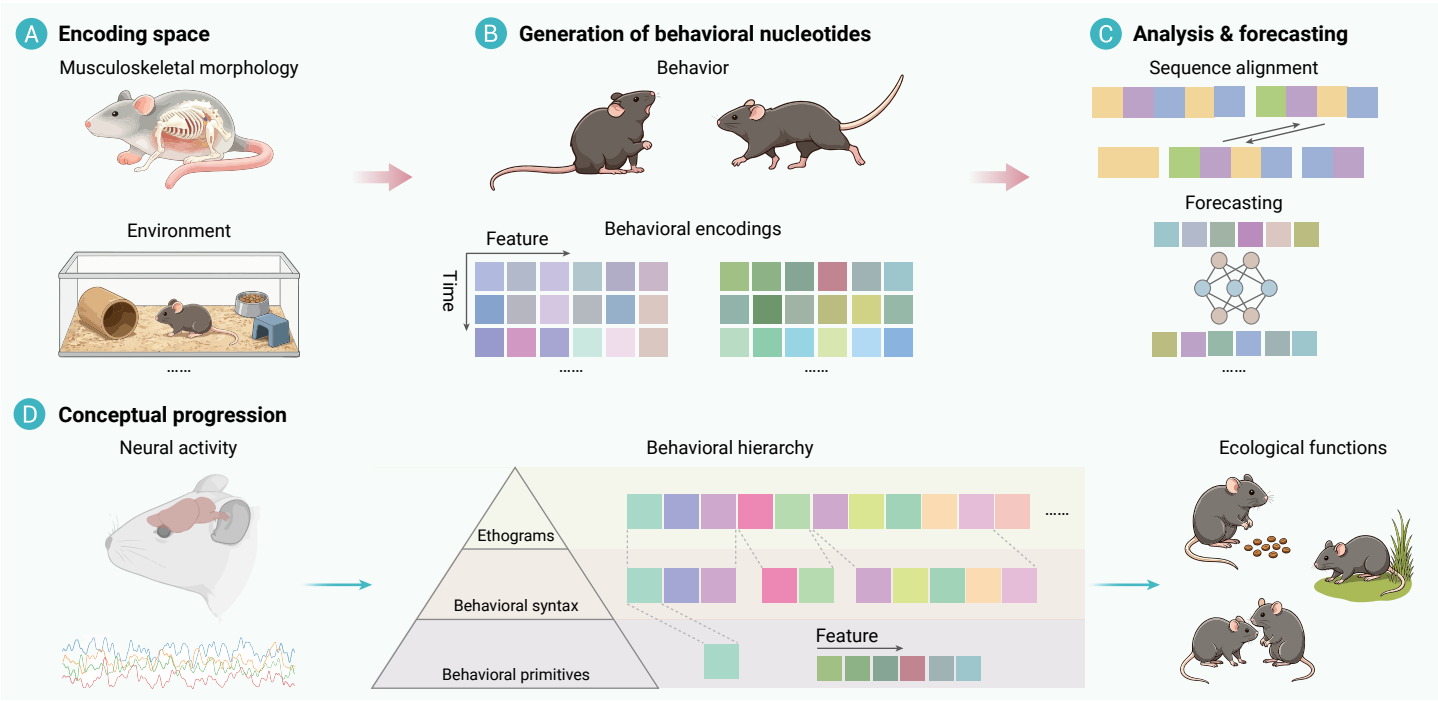


Figure 6. Behavioral nucleotides as a speculative framework for behavioral encoding, sequence analysis, and hierarchical organization (A) Encoding space. Animal morphology, biomechanics, and environmental context constrain the possible forms of behavioral nucleotides. (B) Generation of behavioral nucleotides. Observable behavior is sampled and represented as time-dependent multidimensional feature vectors. In the matrices, rows denote successive sampled time points and columns denote feature dimensions; each sampled time point, represented by a multidimensional feature vector, can be treated as a candidate behavioral nucleotide. (C) Analysis and forecasting. The resulting behavioral representations can support sequence alignment and probabilistic forecasting of future behavioral encodings. (D) Conceptual progression. A possible relationship is illustrated from neural activity to a hierarchical behavioral organization. Candidate behavioral nucleotides can form ordered sequences or patterns through context-dependent transition and combination rules, here termed behavioral syntax; these higher-order structures may form ethograms linked to ecological functions such as foraging, social interaction, and avoidance or survival.

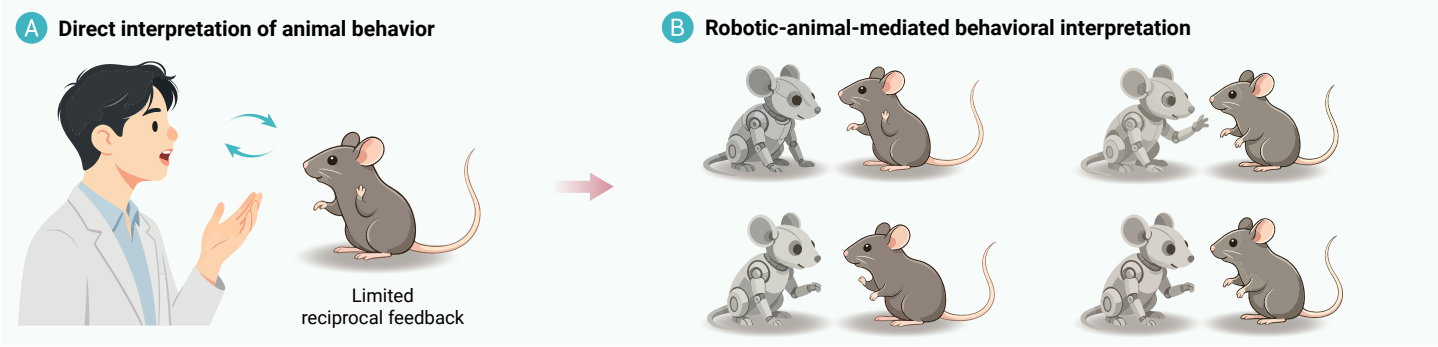


Figure 7. Exploratory use of "bilingual robotic animals" (A) Direct interpretation of animal behavior. Cross-species interpretation is constrained by the absence of a shared semantic feedback channel, resulting in limited reciprocal feedback. (B) Robotic-animal-mediated behavioral interpretation. Controllable robotic animals can emit standardized behavioral signals to test context-dependent relationships between specific signals and observable animal responses.

sitions and temporal variations in movement. Multiscale temporal representations, continuous latent features, and probabilistic encoding may help preserve such information while retaining the advantages of sequence-level analysis.

At a higher organizational level, candidate behavioral nucleotides would combine through context-dependent transition and combination rules, here termed behavioral syntax, to form structured ethograms linked to ecological functions (Figure 6D). The usefulness of such a representation could be assessed through behavioral reconstruction, correspondence with neural activity, causal perturbation where feasible, and generalization across individuals and contexts.

From animal body language to human language

Translating animal body language into human-understandable language is a challenging cross-species semantic-conversion task. Human communication can be checked through reciprocal and mutually interpretable feedback, including verbal confirmation and conventional gestures.[98] Cross-species interpretations, by contrast, cannot be validated through an equivalent shared semantic channel (Figure 7A).

One possible pathway involves utilizing "bilingual robotic animals" as research tools to investigate the functions and correspondences of behavioral signals[78,79,99] (Figure 7B). The principal methodological advantage of robotic animals lies in their capacity to emit standardized, repeatable, and precisely controllable behavioral signals or sequences.[100] Researchers can design robotic animals to exhibit specific postures, action sequences, or combinatorial signals and then place them in social interaction scenarios with live animals for controlled experiments and repeated validation. By precisely manipulating signal variables such as intensity, frequency, duration, or combinatorial pattern and observing changes in the live animal's response, researchers may identify which behavioral features contribute to effective social signals and which responses they elicit in specific contexts.

Ecological validity is an important consideration for robotic-animal experiments. Differences in morphology, movement, acoustic or olfactory cues may alter the responses of live animals, and observed effects may also reflect novelty or general arousal. Comparisons with live conspecifics or biologically realistic stimuli, together with appropriate control conditions, can therefore

help distinguish responses to specific behavioral signals from nonspecific effects.

These controlled signal–response experiments provide an operational entry point for investigating conserved behavioral signals within social interactions and may gradually support the construction of a preliminary translation dictionary.[100] If sufficiently reliable signal–response correspondences can be established, they may eventually support AI-assisted interpretation of animal communication and provide a basis for investigating its underlying neural mechanisms and social functions.

CONCLUSION

The integration of AI into animal behavior research marks a paradigm shift, transitioning the field from traditional qualitative descriptions and limited quantitative measures to high-throughput, objective, and deeply nuanced analyses. As reviewed herein, AI has furnished powerful tools for dissecting behavior at multiple levels, ranging from precise pose estimation and 3D reconstruction to sophisticated classification, individual identification, and social communication. MM-LLMs may help integrate diverse data streams for higher-level behavioral interpretation, while AI agents may coordinate specialized analytical tools across multistep workflows. Nevertheless, the path toward a comprehensive AI-driven ethology is not without significant hurdles; challenges such as data heterogeneity, the subjectivity of behavioral definitions, the combinatorial complexity of social interactions, data acquisition in naturalistic environments, and the robust decoding of intricate behavioral sequences and individual tracking remain formidable.

Continued advances in AI for ethology will depend on rigorous validation, interoperable data standards, and realistic assessment of model limitations. Standardized and physically grounded posture models, virtual environments for training and evaluation, and the forward-looking concepts of "behavioral nucleotides" and "bilingual robotic animals" may provide useful directions for future research.

AI-based behavioral phenotyping may strengthen preclinical research by improving measurement resolution, reproducibility, and the detection of treatment-related behavioral changes. However, translation from animal behavior to human clinical benefit remains constrained by species differences, model validity, and the limited correspondence between laboratory assays and human neuropsychiatric conditions.

In conclusion, AI is reshaping ethology by enabling more detailed, scalable, and reproducible measurement of animal behavior. Continued progress will require careful integration of computational advances with biological validation, transparent benchmarks, and expert interpretation. These efforts can deepen understanding of animal behavior and strengthen the measurement foundations of translational research.

DECLARATION OF AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT solely for language polishing. The authors subsequently reviewed and edited the manuscript as needed and take full responsibility for its content.

REFERENCES

1. Tinbergen N. (1963). On aims and methods of ethology. *Z. Tierpsychol.* **20**:410–433. DOI:10.1111/j.1439-0310.1963.tb01161.x
2. Datta S. R., Anderson D. J., Branson K., et al. (2019). Computational neuroethology: a call to action. *Neuron* **104**:11–24. DOI:10.1016/j.neuron.2019.09.038
3. Miller C. T., Gire D., Hoke K., et al. (2022). Natural behavior is the language of the brain. *Curr. Biol.* **32**:R482–R493. DOI:10.1016/j.cub.2022.03.031
4. Chen Y., Lin Y.C., Kuo T.W., et al. (2015). Sensory detection of food rapidly modulates arcuate feeding circuits. *Cell* **160**:829–841. DOI:10.1016/j.cell.2015.01.033
5. Yilmaz M. and Meister M. (2013). Rapid innate defensive responses of mice to looming visual stimuli. *Curr. Biol.* **23**:2011–2015. DOI:10.1016/j.cub.2013.08.015
6. Wei P., Liu N., Zhang Z., et al. (2015). Processing of visually evoked innate fear by a non-canonical thalamic pathway. *Nat. Commun.* **6**:6756. DOI:10.1038/ncomms7756
7. Krakauer J. W., Ghazanfar A. A., Gomez-Marin A., et al. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**:480–490. DOI:10.1016/j.neuron.2016.12.041
8. Wiltschko A. B., Johnson M. J., Iurilli G., et al. (2015). Mapping sub-second structure in mouse behavior. *Neuron* **88**:1121–1135. DOI:10.1016/j.neuron.2015.11.031
9. Pereira T. D., Shaevitz J. W. and Murthy M. (2020). Quantifying behavior to understand the brain. *Nat. Neurosci.* **23**:1537–1549. DOI:10.1038/s41593-020-00734-z
10. Huang K., Han Y., Chen K., et al. (2021). A hierarchical 3D-motion learning framework for animal spontaneous behavior mapping. *Nat. Commun.* **12**:2784. DOI:10.1038/s41467-021-22970-y
11. Anderson D. J. and Perona P. (2014). Toward a science of computational ethology. *Neuron* **84**:18–31. DOI:10.1016/j.neuron.2014.09.005
12. Gris K.V., Coutu J.P. and Gris D. (2017). Supervised and unsupervised learning technology in the study of rodent behavior. *Front. Behav. Neurosci.* **11**:141. DOI:10.3389/fnbeh.2017.00141
13. Dellu F., Mayo W., Cherkaoui J., et al. (1992). A two-trial memory task with automated recording: study in young and aged rats. *Brain Res.* **588**:132–139. DOI:10.1016/0006-8993(92)91352-F
14. Moy S. S., Nadler J. J., Perez A., et al. (2004). Sociability and preference for social novelty in five inbred strains: an approach to assess autistic-like behavior in mice. *Genes Brain Behav.* **3**:287–302. DOI:10.1111/j.1601-1848.2004.00076.x
15. LeCun Y., Bengio Y. and Hinton G. (2015). Deep learning. *Nature* **521**:436–444. DOI:10.1038/nature14539
16. Mathis A., Mamidanna P., Cury K. M., et al. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**:1281–1289. DOI:10.1038/s41593-018-0209-y
17. Nath T., Mathis A., Chen A. C., et al. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14**:2152–2176. DOI:10.1038/s41596-019-0176-0
18. Berman G. J., Choi D. M., Bialek W., et al. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**:20140672. DOI:10.1098/rsif.2014.0672
19. Han Y., Chen K., Wang Y., et al. (2024). Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nat. Mach. Intell.* **6**:48–61. DOI:10.1038/s42256-023-00776-5
20. Hsu A. I. and Yttri E. A. (2021). B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nat. Commun.* **12**:5188. DOI:10.1038/s41467-021-25420-x
21. Segalin C., Williams J., Karigo T., et al. (2021). The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* **10**:e63720. DOI:10.7554/eLife.63720
22. de Chaumont F., Ey E., Torquet N., et al. (2019). Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. *Nat. Biomed. Eng.* **3**:930–942. DOI:10.1038/s41551-019-0396-1
23. Vaswani A., Shazeer N., Parmar N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, pp:5998-6008. DOI:10.5555/3295222.3295349.
24. Ye S., Lauer J., Zhou M., et al. (2023). AmadeusGPT: a natural language interface for interactive animal behavioral analysis. Advances in Neural Information Processing Systems, pp:6297-6329. DOI:10.52202/075280-0276.
25. Xu T., Zhou T., Wang Y., et al. (2025). MouseGPT: a large-scale vision-language model for mouse behavior analysis. *bioRxiv*. DOI:10.1101/2025.03.27.645630.
26. Hurst J. L. and West R. S. (2010). Taming anxiety in laboratory mice. *Nat. Methods* **7**:825–826. DOI:10.1038/nmeth.1500
27. Ho H., Kejzar N., Sasaguri H., et al. (2023). A fully automated home cage for long-term continuous phenotyping of mouse cognition and behavior. *Cell Rep. Methods* **3**:100532. DOI:10.1016/j.crmeth.2023.100532
28. Kahnau P., Mieske P., Wilzopolski J., et al. (2023). A systematic review of the development and application of home cage monitoring in laboratory mice and rats. *BMC Biol.* **21**:256. DOI:10.1186/s12915-023-01751-7
29. Pereira T. D., Aldarondo D. E., Willmore L., et al. (2019). Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**:117–125. DOI:10.1038/s41592-018-0234-5
30. Lv S., Wang J., Chen X., et al. (2024). STPoseNet: a real-time spatiotemporal network model for robust mouse pose estimation. *iScience* **27**:109772. DOI:10.1016/j.isci.2024.109772
31. Pereira T. D., Tabris N., Matsliah A., et al. (2022). SLEAP: a deep learning system for multi-animal pose tracking. *Nat. Methods* **19**:486–495. DOI:10.1038/s41592-022-01426-1
32. Cao Z., Simon T., Wei S. E., et al. (2017). Realtime multi-person 2D pose estimation using Part Affinity Fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 1302-1310. DOI:10.1109/CVPR.2017.143.
33. Lauer J., Zhou M., Ye S., et al. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* **19**:496–504. DOI:10.1038/s41592-022-01443-0
34. Tang G., Han Y., Sun X., et al. (2025). Anti-drift pose tracker (ADPT), a transformer-based network for robust animal pose estimation cross-species. *eLife* **13**:RP95709. DOI:10.7554/eLife.95709.3
35. Chen Z., Zhang R., Fang H.S., et al. (2023). AlphaTracker: a multi-animal tracking and

behavioral analysis tool. *Front. Behav. Neurosci.* **17**:1111908. DOI:10.3389/fnbeh.2023.1111908
36. Marks M., Jin Q., Sturman O., et al. (2022). Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nat. Mach. Intell.* **4**:331–340. DOI:10.1038/s42256-022-00477-5
37. Blau A., Gebhardt C., Bendesky A., et al. (2022). SemiMultiPose: a semi-supervised multi-animal pose estimation framework. *arXiv*. DOI:10.48550/arXiv.2204.07072.
38. Biderman D., Whiteway M. R., Hurwitz C., et al. (2024). Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling and cloud-native open-source tools. *Nat. Methods* **21**:1316–1328. DOI:10.1038/s41592-024-02319-1
39. Ye S., Filippova A., Lauer J., et al. (2024). SuperAnimal pretrained pose estimation models for behavioral analysis. *Nat. Commun.* **15**:5165. DOI:10.1038/s41467-024-48792-2
40. Li C. and Lee G. H. (2021). From synthetic to real: unsupervised domain adaptation for animal pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp:1482-1491. DOI:10.1109/CVPR46437.2021.00153.
41. Yu H., Xu Y., Zhang J., et al. (2021). AP-10K: a benchmark for animal pose estimation in the wild. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. https://openreview.net/forum?id=rH8yliN6C83.
42. Marshall J. D., Klibaite U., Gellis A., et al. (2021). The PAIR-R24M dataset for multi-animal 3D pose estimation. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. https://openreview.net/forum?id=-wVVl_UPr8.
43. Ebbesen C. L. and Froemke R. C. (2022). Automatic mapping of multiplexed social receptive fields by deep learning and GPU-accelerated 3D videography. *Nat. Commun.* **13**:593. DOI:10.1038/s41467-022-28153-7
44. Zuffi S., Kanazawa A., Jacobs D. W., et al. (2017). 3D Menagerie: modeling the 3D shape and pose of animals. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp:5524-5532. DOI:10.1109/CVPR.2017.586.
45. Han Y., Huang K., Chen K., et al. (2022). MouseVenue3D: a markerless three-dimension behavioral tracking system for matching two-photon brain imaging in free-moving mice. *Neurosci. Bull.* **38**:303–317. DOI:10.1007/s12264-021-00778-6
46. Dunn T. W., Marshall J. D., Severson K. S., et al. (2021). Geometric deep learning enables 3D kinematic profiling across species and environments. *Nat. Methods* **18**:564–573. DOI:10.1038/s41592-021-01106-6
47. Klibaite U., Li T., Aldarondo D., et al. (2025). Mapping the landscape of social behavior. *Cell* **188**:2249–2266.e23. DOI:10.1016/j.cell.2025.01.044
48. Stern U., He R. and Yang C.H. (2015). Analyzing animal behavior via classifying each video frame using convolutional neural networks. *Sci. Rep.* **5**:14351. DOI:10.1038/srep14351
49. Tang C., Zhou Y., Zhao S., et al. (2024). Segmentation tracking and clustering system enables accurate multi-animal tracking of social behaviors. *Patterns* **5**:101057. DOI:10.1016/j.patter.2024.101057
50. Weinreb C., Pearl J. E., Lin S., et al. (2024). Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nat. Methods* **21**:1329–1339. DOI:10.1038/s41592-024-02318-2
51. von Ziegler L., Sturman O. and Bohacek J. (2021). Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacology* **46**:33–44. DOI:10.1038/s41386-020-0751-7
52. Markowitz J. E., Gillis W. F., Beron C. C., et al. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. *Cell* **174**:44–58.e17. DOI:10.1016/j.cell.2018.04.019
53. Goodwin N. L., Choong J. J., Hwang S., et al. (2024). Simple Behavioral Analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nat. Neurosci.* **27**:1411–1424. DOI:10.1038/s41593-024-01649-9
54. Schneider S., Lee J. H. and Mathis M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**:360–368. DOI:10.1038/s41586-023-06031-6
55. Pereira E., Araújo Í. and Silva L.F.V. (2023). RFID technology for animal tracking: a survey. *IEEE J. Radio Freq. Identif.* **7**:609–620. DOI:10.1109/JRFID.2023.3334952
56. Pérez-Escudero A., Vicente-Page J., Hinz R. C., et al. (2014). idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nat. Methods* **11**:743–748. DOI:10.1038/nmeth.2994
57. Romero-Ferrero F., Bergomi M. G., Hinz R. C., et al. (2019). idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nat. Methods* **16**:179–182. DOI:10.1038/s41592-018-0295-5
58. Robinson I., Glidden-Handgis G., Panchal N., et al. (2025). DIPLOMAT: multi-animal tracking with efficient manual editing. *bioRxiv*. DOI:10.1101/2025.08.11.669786.
59. Torrents J., Costa T. and de Polavieja G. (2026). New idtracker.ai rethinks multi-animal tracking as a representation learning problem to increase accuracy and reduce tracking time. *eLife* **14**:RP107602. DOI:10.7554/eLife.107602.4
60. Li Y., Zhang Q., Zhang Y., et al. (2026). Unsupervised transfer learning enables multi-animal tracking without training annotation. *Nat. Methods* **23**:1058–1068. DOI:10.1038/s41592-026-03051-8
61. Yin S., Fu C., Zhao S., et al. (2024). A survey on multimodal large language models. *Natl. Sci. Rev.* **11**:nwae403. DOI:10.1093/nsr/nwae403
62. Xi Z., Chen W., Guo X., et al. (2025). The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.* **68**:121101. DOI:10.1007/s11432-024-4222-0
63. Wang Y., Li X., Yan Z., et al. (2025). InternVideo2.5: empowering video MLLMs with long and rich context modeling. *arXiv*. DOI:10.48550/arXiv.2501.12386
64. Bai S., Cai Y., Chen R., et al. (2025). Qwen3-VL technical report. *arXiv*. DOI:10.48550/arXiv.2511.21631
65. Crabbe J. C., Wahlsten D. and Dudek B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**:1670–1672. DOI:10.1126/science.284.5420.1670
66. Qwen Team. (2026). Qwen3.5-Omni technical report. *arXiv*. DOI: 10.48550/arXiv.2604.15804
67. Coffey K. R., Marx R. E. and Neumaier J. F. (2019). DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology* **44**:859–868. DOI:10.1038/s41386-018-0303-6
68. Aljović A., Lin Z., Wang W., et al. (2025). An autonomous AI agent for universal behavior analysis. *bioRxiv*. DOI:10.1101/2025.05.15.653585
69. Zhou M., Zhou S., Xiong B., et al. (2026). Humanoid animal behavior labeler: an interactive annotation agent to accelerate ethological applications in precision management of animals using large vision-language models. *Comput. Electron. Agric.* **241**:111307. DOI:10.1016/j.compag.2025.111307
70. Chen K., Chen Z., Zheng D., et al. (2026). EthoClaw: an integrated AI workflow platform for automated analysis in neuroethology. *bioRxiv*. DOI:10.64898/2026.03.25.714141
71. Yang Y., Yang J., Xu Y., et al. (2022). APT-36K: a large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems* **35**:17301–17313. DOI:10.52202/068431-1258
72. Sun J. J., Karigo T., Chakraborty D., et al. (2021). The Multi-Agent Behavior Dataset: mouse dyadic social interactions. *In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* **1**:1–15.
73. Zhang L., Gao J., Xiao Z., et al. (2023). AnimalTrack: a benchmark for multi-animal tracking in the wild. *Int. J. Comput. Vis.* **131**:496–513. DOI:10.1007/s11263-022-01711-8
74. Duty S. and Jenner P. (2011). Animal models of Parkinson's disease: a source of novel treatments and clues to the cause of the disease. *Br. J. Pharmacol.* **164**:1357–1391. DOI:10.1111/j.1476-5381.2011.01426.x
75. Rübel O., Tritt A., Ly R., et al. (2022). The Neurodata Without Borders ecosystem for neurophysiological data science. *eLife* **11**:e78362. DOI:10.7554/eLife.78362
76. Dodge S. and Karam L. (2016). Understanding how image quality affects deep neural networks. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp:1-6. DOI:10.1109/QoMEX.2016.7498955
77. Ebbesen C. L. and Froemke R. C. (2021). Body language signals for rodent social communication. *Curr. Opin. Neurobiol.* **68**:91–106. DOI:10.1016/j.conb.2021.01.008
78. Quinn L. K., Schuster L. P., Aguilar-Rivera M., et al. (2018). When rats rescue robots. *Anim. Behav. Cogn.* **5**:368–379. DOI:10.26451/abc.05.04.04.2018
79. Jia G., Chen Z., Zhang Y., et al. (2024). Modulating emotional states of rats through a rat-like robot with learned interaction patterns. *Nat. Mach. Intell.* **6**:1580–1593. DOI:10.1038/s42256-024-00939-y
80. Han Y., Jiang Z., Ju F., et al. (2025). BL-BERT: extracting body language from behavior sequences in freely moving mice. Human Brain and Artificial Intelligence: 4th International Workshop, HBAI 2024, Proceedings (Springer), pp:181-191. DOI:10.1007/978-981-96-4001-0_12
81. Wynne C. D. L. (2007). What are animals? Why anthropomorphism is still not a scientific approach to behavior. *Comp. Cogn. Behav. Rev.* **2**:125–135. DOI:10.3819/CCBR.2008.20008
82. Botvinick M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* **12**:201–208. DOI:10.1016/j.tics.2008.02.009
83. Seyfarth R. M., Cheney D. L. and Marler P. (1980). Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Anim. Behav.* **28**:1070–1094. DOI:10.1016/S0003-3472(80)80097-2
84. Festing M. F. W. and Altman D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J.* **43**:244–258. DOI:10.1093/ilar.43.4.244
85. Mathis M. W. and Mathis A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **60**:1–11. DOI:10.1016/j.conb.2019.10.008
86. Galoaa B., Amraee S. and Ostadabbas S. (2025). More than meets the eye: enhancing multi-object tracking even with prolonged occlusions. In Proceedings of the 42nd International Conference on Machine Learning(PMLR), pp:18132-18142.
87. Bolaños L. A., Xiao D., Ford N. L., et al. (2021). A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nat. Methods* **18**:378–381. DOI:10.1038/s41592-021-01103-9
88. Aldarondo D., Merel J., Marshall J. D., et al. (2024). A virtual rodent predicts the

structure of neural activity across behaviours. *Nature* **632**:594–602. DOI:10.1038/s41586-024-07633-4
89. Zhang C. Y., Yang Y., Sirbu A., et al. (2025). MIMIC-MJX: neuromechanical emulation of animal behavior. *arXiv*. DOI:10.48550/arXiv.2511.20532
90. National Institutes of Health. (2025). NIH to prioritize human-based research technologies. https://www.nih.gov/news-events/news-releases/nih-prioritize-human-based-research-technologies
91. Dosovitskiy A., Ros G., Codevilla F., et al. (2017). CARLA: an open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning(PMLR), pp:1-16.
92. Makoviychuk V., Wawrzyniak L., Guo Y., et al. (2021). Isaac Gym: high performance GPU based physics simulation for robot learning. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/28dd2c7955ce926456240b2ff0100bde-Abstract-round2.html
93. DeWolf T., Schneider S., Soubiran P., et al. (2024). Neuro-musculoskeletal modeling reveals muscle-level neural dynamics of adaptive learning in sensorimotor cortex. *bioRxiv*. DOI:10.1101/2024.09.11.612513
94. Simon H. A. (1962). The architecture of complexity. *Proc. Am. Philos. Soc.* **106**:467–482.
95. Altschul S. F., Gish W., Miller W., et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410. DOI:10.1016/S0022-2836(05)80360-2
96. Bailey T. and Elkan C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**:28–36.
97. Avsec Ž., Agarwal V., Visentin D., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**:1196–1203. DOI:10.1038/s41592-021-01252-x
98. Clark H. H. and Brennan S. E. (1991). Grounding in communication. Resnick L. B., Levine J. M. and Teasley S. D. (eds). Perspectives on socially shared cognition (American Psychological Association), pp: 127-149. DOI: 10.1037/10096-006
99. Halloy J., Sempo G., Caprari G., et al. (2007). Social integration of robots into groups of cockroaches to control self-organized choices. *Science* **318**:1155–1158. DOI:10.1126/science.1144259
100. Krause J., Winfield A. F. T. and Deneubourg J.L. (2011). Interactive robots in experimental biology. *Trends Ecol. Evol.* **26**:369–375. DOI:10.1016/j.tree.2011.03.015

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the STI2030-Major Projects (2021ZD0203900), National Natural Science Foundation of China (T2394533), Jiangsu Provincial Major Science and Technology Special Project (BG2025040), Start-up Research Fund of Southeast University (RF1028625096). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

K.C., Y.H., and K.H. conceptualized the manuscript. K.C. and Y.H. drafted the initial manuscript. Y.H., K.H., L.W., and P.W. contributed substantially to the revision of the manuscript. K.C. and Y.H. prepared all figures. All authors reviewed and edited the manuscript and approved the final version.

DECLARATION OF INTERESTS

K.C. and K.H. are employees of Guangdong BayOne Biotech Co., Ltd., which holds patents related to the BehaviorAtlas technology that were transferred from the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. L.W. and P.W. hold no equity or other financial interests in Guangdong BayOne Biotech Co., Ltd. The other authors declare no competing interests.

DATA AND CODE AVAILABILITY

No new data were created or analyzed in this study.