nature methods

Article

https://doi.org/10.1038/s41592-026-03051-8

Unsupervised transfer learning enables multi-animal tracking without training annotation

Received: 23 January 2025

Accepted: 27 February 2026

Published online: 04 May 2026

Check for updates

Yixin Li[1,2,3], Qi Zhang[1,4], Yuanlong Zhang[5,6], Jiaqi Fan[1,7], Zhi Lu[8], Xinhong Xu[1,4], Xinyang Li[2] ✉, Ziwei Li[3,9] ✉, Jiamin Wu[1,4,6] ✉ & Qionghai Dai[1,4,6] ✉

Quantitative ethology necessitates accurate tracking of animal locomotion, especially for population-level analyses involving multiple individuals. However, current methods mostly rely on laborious annotations for supervised training and have restricted performance under challenging conditions. Here we present an unsupervised deep transfer learning method for multi-animal tracking (UDMT) that achieves state-of-the-art performance without requiring training annotations. By synergizing a bidirectional closed-loop tracking strategy, a spatiotemporal transformer network and three dedicated modules for localization refining, bidirectional identity correction and automatic parameter tuning, UDMT can track multiple animals accurately under various challenging conditions, such as crowding, occlusion, rapid motion, low image contrast and cross-species experiments. We demonstrate the versatility of UDMT on five different kinds of model animals, including mice, rats, *Drosophila*, *Caenorhabditis elegans* and *Betta splendens*. Combined with a head-mounted miniaturized microscope, we illustrate the power of UDMT for neuroethological interrogations to decipher the correlations between animal locomotion and neural activity.

Animal behavior reflects their internal states and external conditions, which jointly shape how animals move, interact and respond to the environment[1,2]. Quantifying animal behavior is a fundamental step in ethology, neuroscience, psychology and various other fields[3–5]. As the most basic representation of behavior, the position of animals can reflect the locomotion of individuals and serves as a metric for behavioral analysis, especially for population-level studies involving multiple animals. Over the past few decades, the technology for animal tracking has evolved continuously, and recent advances have catalyzed a series of scientific discoveries[6,7]. However, challenges such as similarities in appearance and frequent interactions of animals hinder the development of multi-animal tracking towards higher accuracy, larger scale and more complex scenarios.

Artificial intelligence has been adopted with remarkable success in animal tracking[6,8–12]. Supervised-learning-based algorithms can achieve good performance[10,13,14], but they require manual annotations

[1]Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. [2]College of AI, Tsinghua University, Beijing, China. [3]College of Future Information Technology, Fudan University, Shanghai, China. [4]Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing, China. [5]School of Life Sciences, Tsinghua University, Beijing, China. [6]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China. [7]Institute of Digital Medicine, City University of Hong Kong, Hong Kong, China. [8]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China. [9]Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai, China. ✉e-mail: xinyangli@tsinghua.edu.cn; lizw@fudan.edu.cn; wujiamin@tsinghua.edu.cn; qhdai@tsinghua.edu.cn

for training, which is time-consuming and laborious[15]. To reduce human intervention, straightforward yet effective semi-automatic annotation has been incorporated into animal tracking by TRex[12] and idtracker.ai (IDT.ai)[11], allowing users to interactively select thresholds for animal segmentation through a graphical user interface (GUI). However, threshold segmentation tends to be unreliable in complex environments and low-contrast conditions. Furthermore, in scenarios involving frequent animal interactions and occlusions, existing tracking methods are prone to identity switches (IDSW) due to insufficient mechanisms for correcting anomalous trajectories, resulting in accumulated errors that substantially degrade tracking accuracy.

Unsupervised learning shows potential to eliminate the reliance on human annotation or ground-truth labels by constructing supervisory relationships directly from data, instead of resorting to external labels[16,17]. Unsupervised learning can perform better than supervised methods when applied to test datasets[18,19], providing a feasible methodology for achieving higher accuracy with minimal annotation costs[20–22]. Moreover, unsupervised learning can eliminate annotation bias inherent in supervised methods[23], which is caused by human variability and mistakes, or by insufficient labeling diversity that fails to represent the entire dataset. The emergence of foundation models has opened up opportunities, as these generalizable pretrained models can provide good starting points and scalable auxiliary components for unsupervised training[24,25]. Despite these expectations, unsupervised learning remains underexplored in multi-animal tracking, particularly under challenging conditions of complex backgrounds, low image contrast and frequent occlusions.

Here, we present an unsupervised deep transfer learning method for multi-animal tracking (UDMT) that outperforms existing tracking methods. UDMT does not require human annotations for training. It is based on a bidirectional closed-loop tracking strategy that enables unsupervised network training by imposing consistency between forward and backward tracking. To capture the spatiotemporal evolution of animal features more effectively, we incorporated a spatiotemporal transformer network (ST-Net) to utilize self-attention and cross-attention mechanisms for feature extraction, leading to a threefold reduction in IDSW compared with convolutional neural networks (CNNs). For identity (ID) correction in the inference stage, we designed a dedicated module utilizing backward tracking to relocate missing targets caused by crowding and occlusion, achieving a 2.7-fold improvement in tracking accuracy. We demonstrate the state-of-the-art performance of UDMT on five different kinds of model animals, including mice, rats, *Drosophila*, *Caenorhabditis elegans* and *Betta splendens*. Combined with a head-mounted miniaturized microscope, we recorded the calcium transients synchronized with mouse locomotion to decipher the correlations between animal locomotion and neural activity. We have released the Python source code and a user-friendly GUI of UDMT to make it an accessible tool for quantitative ethology and neuroethology.

Results

Principle of UDMT

The principle of UDMT is illustrated in Fig. 1. The rationale for our unsupervised strategy is that a tracker should be effective in both forward and backward predictions[26]. In other words, the tool should track the target forward in time and subsequently track it backward in time to return to its initial position (Fig. 1a). To establish the internal supervision for network training, we constructed a consistency loss to constrain the deviation between forward tracking and backward tracking. To capture discriminative features, we integrated a spatiotemporal transformer network[27] (ST-Net) to simultaneously extract spatial features of animal appearance and temporal correlations between successive frames (Fig. 1b). The encoder of ST-Net leverages self-attention blocks to aggregate multiple time-variant features of the template (an image patch from a previous frame that contains the target animal), endowing it with the capability to utilize temporal information as the animal posture changes over time. Cross-attention blocks in the decoder bridge the template branch and the search branch to determine the location of the animal inside the search area (a larger image patch from the current frame, centered on the predicted position).

The whole workflow consists of a training process and an inference process. To initialize the training process (Fig. 1c), all animals of the same type in the first frame are segmented semi-automatically by a fine-tuned foundation model[28], with optional manual corrections via a GUI (Extended Data Fig. 1). Subsequently, the video and initial positions are passed to a pretrained model to generate a single-object dataset for training. The pretrained model has been trained on a substantial dataset of public videos and can achieve coarse tracking of animals. The network parameters are then optimized by stochastic gradient descent of the cycle-consistency loss. After training converges, specific representations can be learned from the dataset and memorized in the model. To further improve the tracking accuracy, we designed a localization refining module that utilizes a pretrained object segmentation model[29] to segment animals throughout the video (Supplementary Fig. 1). With our unsupervised transfer training strategy and auxiliary operations, the original video is sufficient for training without any manual annotation.

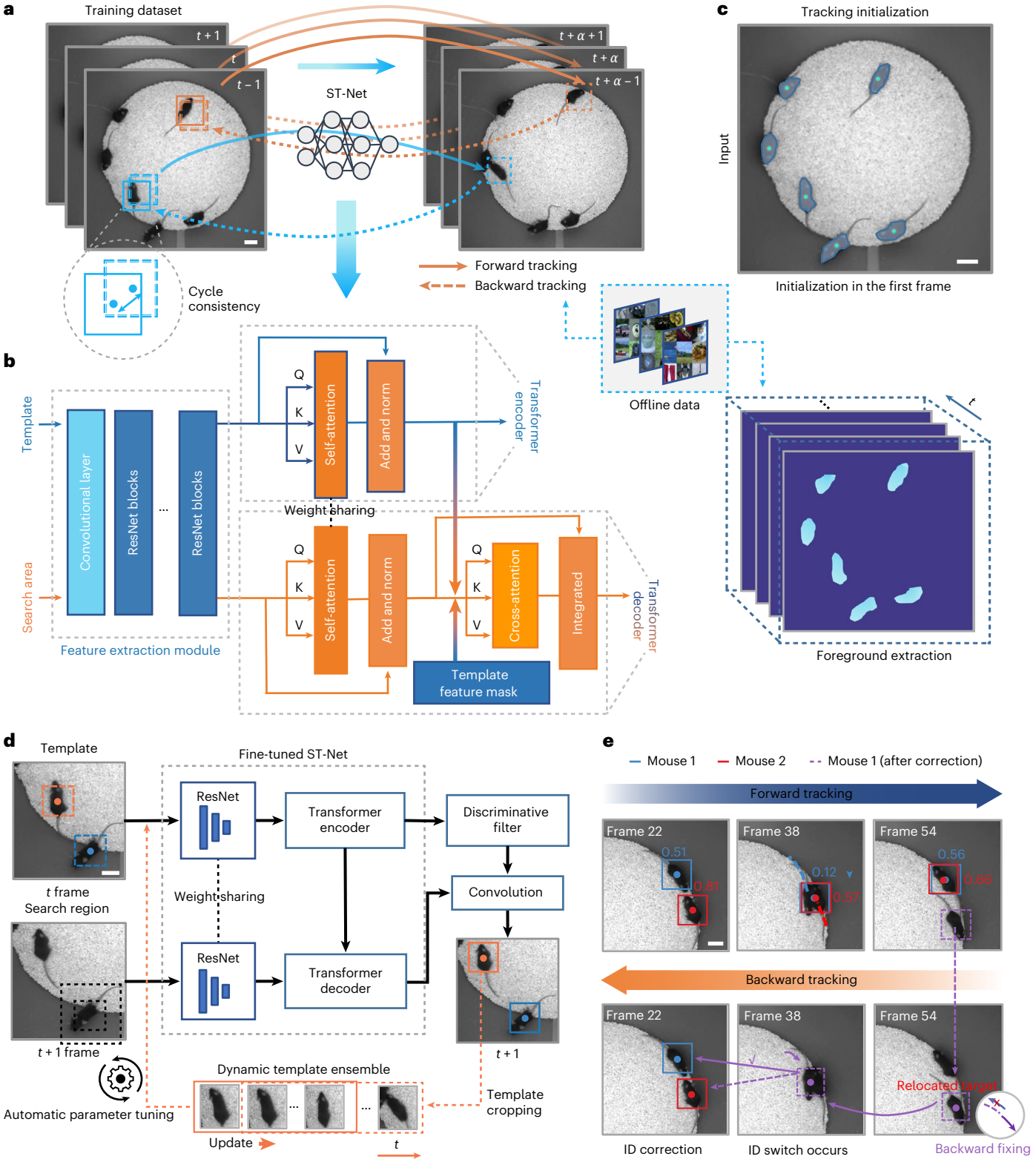
In the inference process, search regions centered on selected animals are cropped out from input frames to eliminate the interference of redundant surrounding pixels. These search regions are then fed into the trained model to infer animal positions in the next frame (Fig. 1d). Predicting the next position using multiple previous frames can utilize the continuity priors of animal locomotion. Since the posture of animals can change over time in behavioral recordings, we developed a method for template updating that can continuously refresh the template ensemble throughout the tracking process to capture dynamic animal features. For deep-learning-based tracking, the size of the search region is a critical hyperparameter that determines

Fig. 1 | Overview of the workflow and modules of UDMT. a, Unsupervised transfer-learning strategy of UDMT. The video is fed into a pretrained ST-Net and the output coarse tracking results are used to construct the dataset for fine-grained training. For each animal, we track it forward and then backward to form a closed loop. The cycle-consistency loss is computed for backpropagation training of the network. **b**, The architecture of the ST-Net. It consists of a convolutional feature extraction module, a transformer encoder module and a transformer decoder module. From left to right: the first-stage CNN takes the template and search area as input for feature extraction. These features are then passed to the following encoder and decoder. The built-in self-attention blocks leverage shared weights to convert the template and search embeddings into the same feature space. In the decoding process, cross-attention blocks are deployed to merge the template and search branch to integrate temporal contexts. The template refers to an image patch cropped from a previous frame containing the target animal. The search area refers to a larger image patch cropped from the current frame, centered on the predicted position. **c**, Tracking initialization. The initial frame is semi-automatically initialized by a fine-tuned animal instance segmentation network, followed by user correction of the segmentation results. **d**, Deployment of the UDMT model. The template frame (t) and search frame ($t+1$) are fed into the fine-tuned ST-Net simultaneously. The encoder is used to extract template features to train discriminative filters and the decoder is used to extract search features. The convolution of discriminative filters and search features can localize the object. During online tracking, the template ensemble will be continuously updated to incorporate adjacent temporal cues and adapt to the changes of target appearance. An automatic parameter tuning module is also proposed for optimal receptive field searching. **e**, Bidirectional tracking for ID correction. Those lost targets in the forward tracking process will be relocated in backward tracking, and the correct ID will be reassigned according to feature similarity. Trajectories will also be refined during backward tracking. Scale bars: 50 mm for all images. In the transformer modules, Q, K and V denote query, key and value, respectively; 'Add and norm' indicates residual addition followed by layer normalization.

https://doi.org/10.1038/s41592-026-03051-8

the spatial extent of the image processed by the network. Its optimal setting is dependent on various factors, including movement speed, animal size, pixel size and recording frame rate, thus necessitating manual adjustment. In general, a larger search region provides more spatial information at the expense of introducing interference in crowded scenes, which can exacerbate IDSW. To obtain the optimized search region, we designed an automatic parameter tuning module based on the observation that the tracking accuracy is closely related to several explicit metrics, including the number of ID corrections, off-target localizations and missing targets (Extended Data Fig. 2a and Supplementary Table 1). Quantitative results show that our parameter tuning module can automatically find the optimal search region size for a specific dataset without human intervention and lead to high tracking accuracy (Extended Data Fig. 2b).

Another critical challenge in multi-animal tracking is ID error owing to frequent animal interaction and occlusion. If an animal is



https://doi.org/10.1038/s41592-026-03051-8

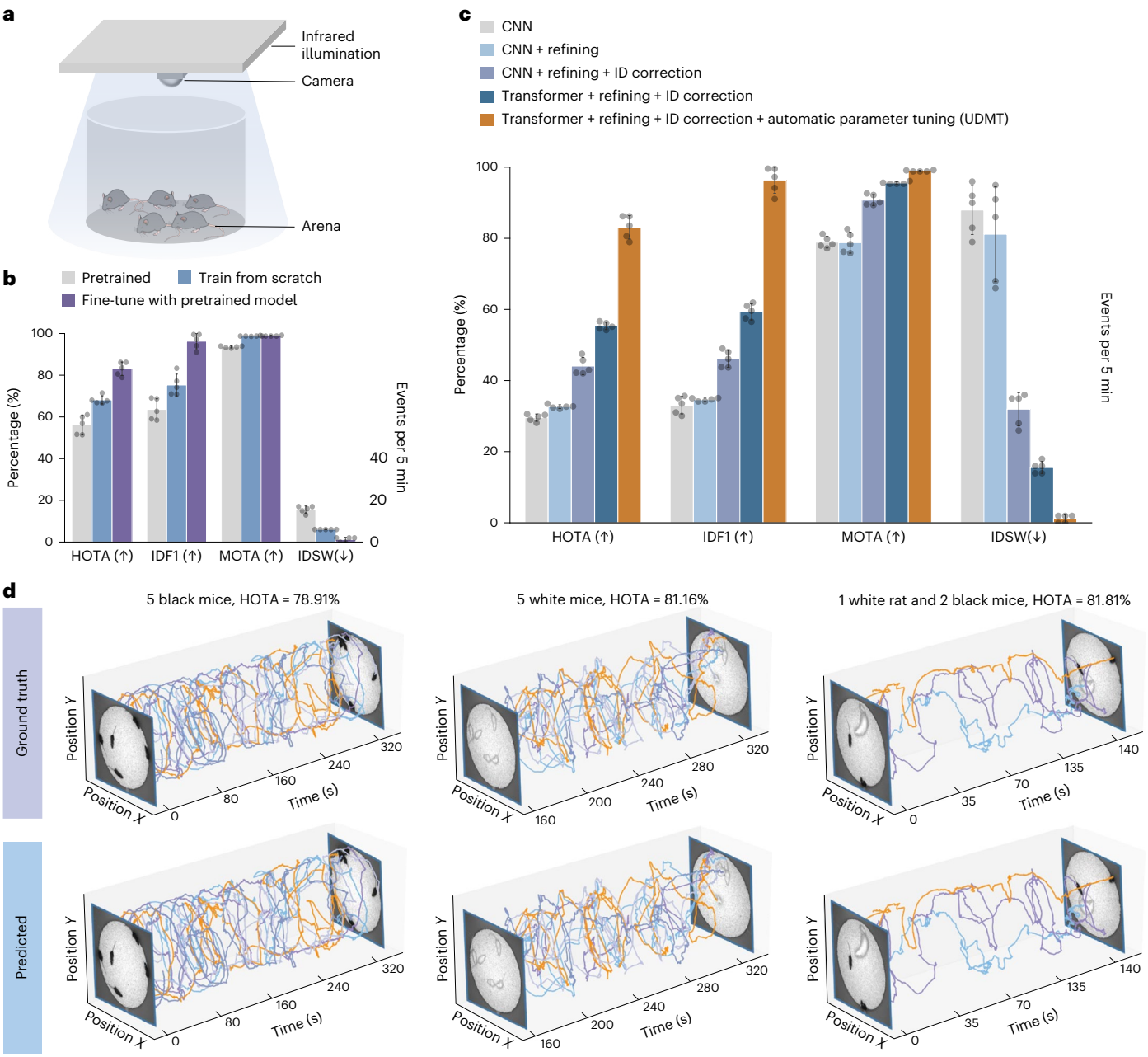


Fig. 2 | Evaluating the performance of UDMT. a, Mouse behavioral recording system. **b**, Tracking performance (quantified by HOTA, MOTA, IDF1 and IDSW) of pretrained models, training from randomly initialized network parameters (training from scratch) and fine-tuning from the pretrained models (unsupervised transfer learning). **c**, Effectiveness of our transformer architecture and the three critical modules (localization refining, ID correction and automatic parameter tuning) in UDMT. **b**,**c**, The 7-mouse dataset (67-Hz frame rate, 29,550 frames) was used for quantitative evaluation. Bars indicate mean values and error bars indicate standard deviation. Gray dots represent individual samples ($n = 5$ video segments). **d**, Manually annotated ground truth and UDMT tracking results of 5 black mice (left), 5 white mice (middle), 1 rat and 2 black mice (right). Different colors represent different individuals.

assigned an incorrect ID, the error propagates throughout subsequent tracking and degrades tracking accuracy. To prevent cumulative ID error, we designed an ID correction module based on bidirectional tracking (Fig. 1e) that can automatically relocate missing targets and reassign correct IDs by detecting abnormal fluctuations of moving speed and localization confidence (Extended Data Fig. 3). After ID correction, missing and misidentified targets in forward tracking will be fixed through backward tracking. In addition to automatic error correction, we integrated a proofreading module into the GUI to eliminate potential errors that cannot be corrected automatically (Extended Data Fig. 1d). Low-confidence frames will be highlighted, allowing users to review and correct remaining errors. Although the improvement brought by proofreading is not included in subsequent evaluations, this module is indispensable as it can compensate for algorithmic limitations and maximize the accuracy of the final results.

We verified the effectiveness of UDMT on top-view behavioral recordings of multiple mice inside a circular arena (Fig. 2a and Supplementary Fig. 2a). For comprehensive comparison of each algorithm configuration, we calculated the Higher Order Tracking Accuracy (HOTA)[30], Multi-Object Tracking Accuracy (MOTA)[31], Identification F1-score (IDF1)[32] and IDSW between predicted trajectories and manually annotated ground truth. These metrics provide a comprehensive evaluation of tracking performance from complementary perspectives. Specifically, HOTA is the primary metric as it balances the

accuracy of object detection and ID association. MOTA emphasizes object detection and is biased towards measuring detection accuracy. IDF1 and IDSW reflect only the accuracy of ID association. To quantify the benefit of unsupervised transfer learning, we compared the performance of the pretrained model, training from randomly initialized network parameters and fine-tuning from the pretrained model (Fig. 2b). The result shows that unsupervised transfer learning leads to a substantial improvement in tracking accuracy, especially in IDSW, which is reduced by 13-fold (1.20 ± 1.10 versus 15.60 ± 1.67, mean ± s.d.) compared with the original pretrained model and fivefold (1.20 ± 1.10 versus 6.00 ± 0.00, mean ± s.d.) compared with training from random initialization. We also conducted a progressive ablation study to assess the contributions of all proposed components (Fig. 2c), confirming that each component not only improves individual aspects of tracking but also works synergistically to achieve state-of-the-art performance. In addition to quantitative metrics, we scrutinized the tracking results by trajectory visualization (Fig. 2d), which indicates that the results of UDMT are consistent with the manually annotated ground truth.

State-of-the-art performance of UDMT under various conditions

The diversity and complexity of animal behavior pose substantial challenges for tracking, including crowding, occlusion, rapid motion and interactions. We tested the effectiveness of UDMT in various tracking scenarios (Fig. 3a), including crowded and occluded black mice (Supplementary Video 1), low-contrast white mice (Supplementary Video 2), cross-species experiments containing one white rat and two black mice (Supplementary Video 3) and sudden rapid motion such as jumping (Supplementary Video 4). With manual annotations as the ground truth, we found that UDMT can handle these challenging conditions without losing or misidentifying animals. The tracking of multiple white mice also reflects that UDMT can tolerate low contrast between animals and the background. Next, we benchmarked UDMT against state-of-the-art methods, including DeepLabCut[14] (DLC), DLC-SuperAnimal[33], SLEAP[10], IDT.ai[11] (IDT.ai) and TRex[12]. We captured and compiled an ethological dataset containing 3–10 mice with recording frame rates ranging from 22 Hz to 94 Hz (Supplementary Table 2). For multi-animal tracking, three dominant factors affect performance: the number of animals, recording duration and frame rate. We started by comparing the accuracy of different methods for tracking different numbers of animals in the same arena (Fig. 3b). Quantitative results show that UDMT outperforms other tracking methods, particularly when the number of mice is larger than five. Specifically, UDMT achieved a HOTA of 71.87 ± 3.20% when tracking ten mice while DLC-SuperAnimal, SLEAP, DLC, TRex and IDT.ai achieved only 56.05 ± 1.17%, 49.71 ± 3.72%, 32.00 ± 4.85%, 25.98 ± 2.74% and 25.52 ± 2.50%, respectively (all values are mean ± s.d.). We mainly attribute this improvement to UDMT's capability to automatically detect and correct IDSW during tracking. We then evaluated the performance of different methods on different recording durations (Fig. 3c). Statistical analysis indicates that UDMT has the best accuracy for both short and long videos. For a long video over 600 s, UDMT achieved a HOTA of 71.46 ± 1.54%, much higher than other methods (IDT.ai 53.87 ± 0.61%, DLC-SuperAnimal 42.25 ± 1.41%, TRex 37.47 ± 0.36%, DLC 34.83 ± 0.69%, SLEAP 34.20 ± 1.02%, mean ± s.d.), demonstrating UDMT's potential for tracking animals during long-term ethological recordings.

Since our method relies on the similarity between two adjacent frames, it tends to be more difficult to track animals in low-frame-rate recordings. We conducted independent repeated experiments to investigate the influence of frame rate on the performance of different methods (Fig. 3d), which shows that UDMT performs better than other methods over a wide range of frame rates (from 22 Hz to 94 Hz). Another advantage of UDMT is the balance between processing speed and tracking accuracy. It can achieve a processing speed of up to 10 frames per second and a HOTA of 80.92 ± 4.72% (mean ± s.d.) at the same time. We also evaluated the performance of these methods on different animal configurations, all showing that our method has the best tracking performance quantified by multiple metrics such as HOTA, MOTA, IDF1 and IDSW (Extended Data Fig. 4). Along with quantitative metrics, we also provide snapshots of the tracking results to visualize the accuracy of animal localization (Fig. 3e). Benefitting from the localization refining module, the prediction of UDMT is more consistent with the ground truth than other methods. Furthermore, we assessed how tracking performance is influenced by image quality. Once trained on a specific recording resolution or brightness, UDMT demonstrated a stable performance advance over other methods across a wide range of image resolution and brightness (Extended Data Fig. 5), making it a valuable tool for diverse applications. It is worth noting that our method has robust tracking performance even at resolution as low as 0.66 mm per pixel, which can be easily achieved by most ethological platforms. Additionally, UDMT is slightly affected by illumination fluctuations (Supplementary Fig. 3) but is sensitive to noise (Supplementary Fig. 4). Therefore, for data with a low signal-to-noise ratio, denoising algorithms[34–36] should be adopted before tracking to mitigate the degradation caused by noise. To assess the scalability of UDMT, we performed extended experiments in complex environments simulating the typical living environment of mice by incorporating food, water sources and shelter-like objects with varied textures, brightness and geometries (Extended Data Fig. 5e). Compared with DLC, DLC-SuperAnimal, SLEAP, IDT.ai and TRex, UDMT achieved superior localization accuracy and ID consistency under such a challenging condition (Extended Data Fig. 5f). These results demonstrate that UDMT maintains robust tracking performance even in visually complex and structured environments.

Neuroethological analysis of multiple freely behaving mice

Deciphering how neural circuits manipulate animal behavior is a major goal in neuroscience. Microscopic interrogation with single-cell resolution in freely behaving animals is a promising technology to gain insights into neural circuits[37,38]. We consider accurate behavioral tracking as important as functional imaging since they jointly provide a complete chain for understanding the correlations between neural activity and animal behavior[39]. For high-accuracy neuroethological analysis of multiple freely behaving mice, we combined UDMT with head-mounted miniaturized microscopy to investigate the influence of spatial position and velocity on neural circuits. We recorded the naturalistic behavior of multiple mice (Fig. 4a), one of which was equipped with a head-mounted miniaturized wide-field microscope weighing 2.5 g (ref. 40). This miniaturized microscope can record neural activity across a 3.6 × 3.6-mm^2 field of view (FOV) at 4-μm lateral resolution. The combination of the miniaturized microscope and an infrared camera above the behavioral arena allows us to capture calcium transients of large-scale neural ensembles synchronized with mouse locomotion and interaction.

We used five transgenic mice (one equipped with the miniaturized microscope and four without) expressing the genetically encoded GCaMP6f calcium indicator[41] specifically in layer 2/3 neurons. Throughout the recording duration, the miniaturized microscope can capture calcium dynamics of more than 2,000 neurons in the primary visual cortex. To associate the neural activity with mouse locomotion, we utilized UDMT to track all five mice and extract their trajectories (Fig. 4b and Supplementary Video 5). Visual inspection suggests obvious variations in the movement trajectory of the mouse equipped with the miniaturized microscope. To quantify the difference, we calculated the accumulated moving distance and instantaneous velocity of each mouse, and analyzed the difference in moving patterns between the mouse with the miniaturized microscope and the other mice without the miniaturized microscope (Fig. 4c). Quantitative analysis showed that the mouse with the miniaturized microscope exhibited reduced moving distance and lower velocity compared with the mice without

https://doi.org/10.1038/s41592-026-03051-8

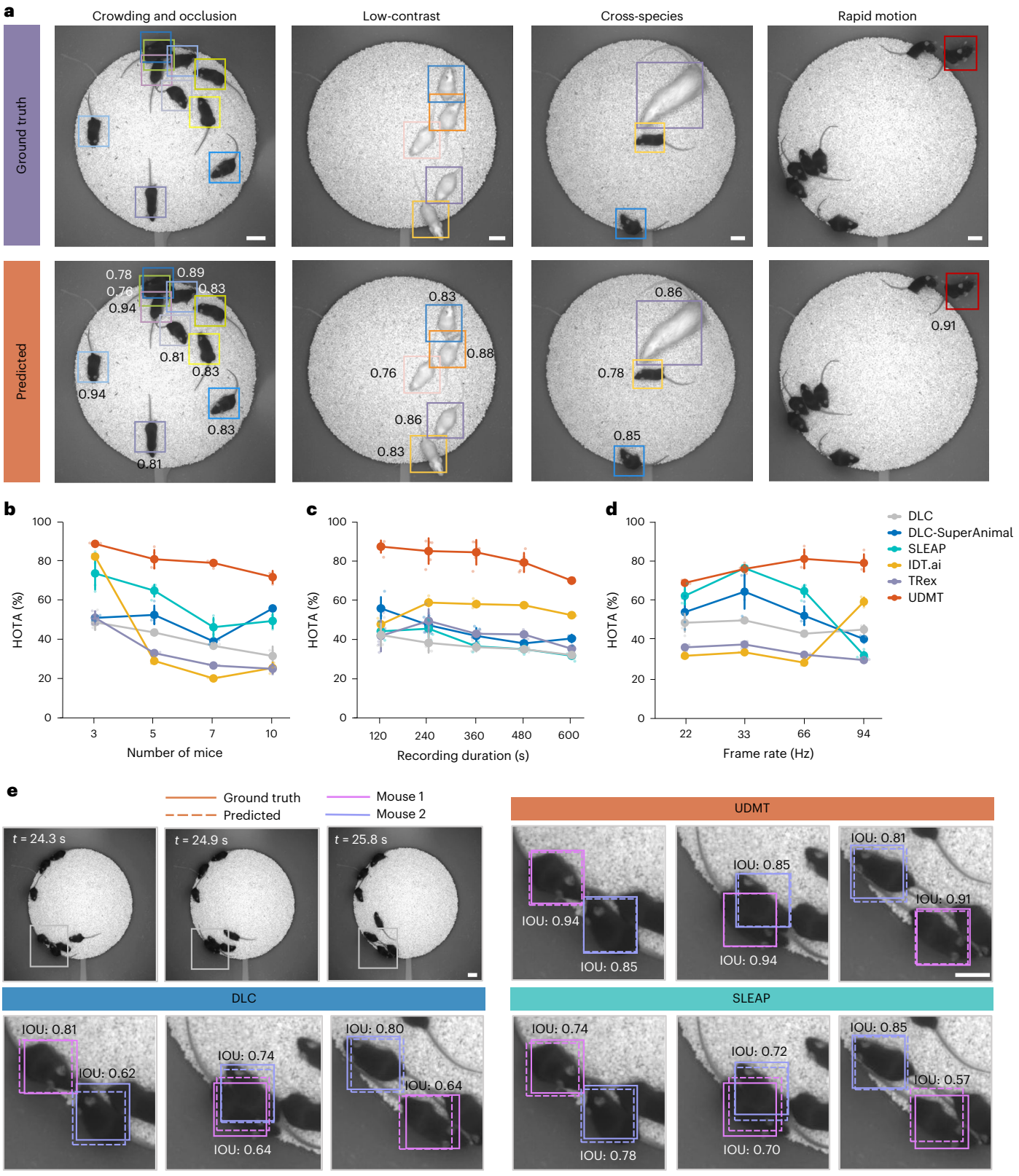


Fig. 3 | State-of-the-art tracking performance of UDMT under various conditions. a, Qualitative evaluation of UDMT under various recording conditions. Numbers near the boxes represent the IOU between the predicted and ground-truth bounding box of target animals. **b**, The relationship between the number of mice and tracking accuracy of different tracking methods. **c**, The relationship between recording duration and tracking accuracy of different tracking methods. **d**, The relationship between recording frame rate and tracking accuracy of different tracking methods. **b**–**d**, Lines indicate mean values and error bars indicate standard deviation. Translucent dots represent individual samples ($n = 5$ video segments per dataset). Datasets included are the 3-, 5-, 7- and 10-mouse datasets (Supplementary Table 2). **e**, Example frames and magnified views of a 7-mouse video at three different time points. Tracking results and corresponding IOU metrics of UDMT, DLC and SLEAP are shown. Scale bars: 50 mm.

miniaturized microscopes (Fig. 4d). Such a reduction in mobility is caused by the weight of the microscope, as well as the weight and tension of the signal transmission cables[42].

We hypothesized that the aggregation state of mice may influence neural activity. We therefore visualized the distance between the mouse with the miniaturized microscope and the other four mice throughout the recording (Fig. 4e). The mean distance is the average distance from the mouse with the microscope to all other mice, and the minimum distance represents the distance from the mouse with the microscope to the nearest mouse. After acquiring a calcium imaging dataset of 2,131 neurons over a 5-min recording period, we quantitatively analyzed the correlations between neural activity and the relative position of the mice (Fig. 4f). We find that the overall spike rate is statistically related to the distance of the mouse with the miniature microscope from its companions. When companions are nearby, the spike rate of observed neurons tends to increase (Fig. 4g). In addition to distance, the first-order differential of spatial position, the velocity, is also significantly correlated to the spike rate of neurons (Fig. 4h; one-sided paired *t*-test, $P = 1.2 \times 10^{-4}$). When the mouse undergoes rapid locomotion, its spike rate tends to increase. We also identified five upregulated neurons that are sensitive to moving velocity (Fig. 4i). Neural activity was different when the mouse with the microscope was near different individuals, suggesting ID-specific effects (Supplementary Fig. 5). In brief, the tracking capability of UDMT facilitates neuroethological analysis and reveals that neurons in the primary visual cortex tend to be more active when the mouse is surrounded by conspecifics or in high-speed locomotion.

Tracking various model animals with UDMT

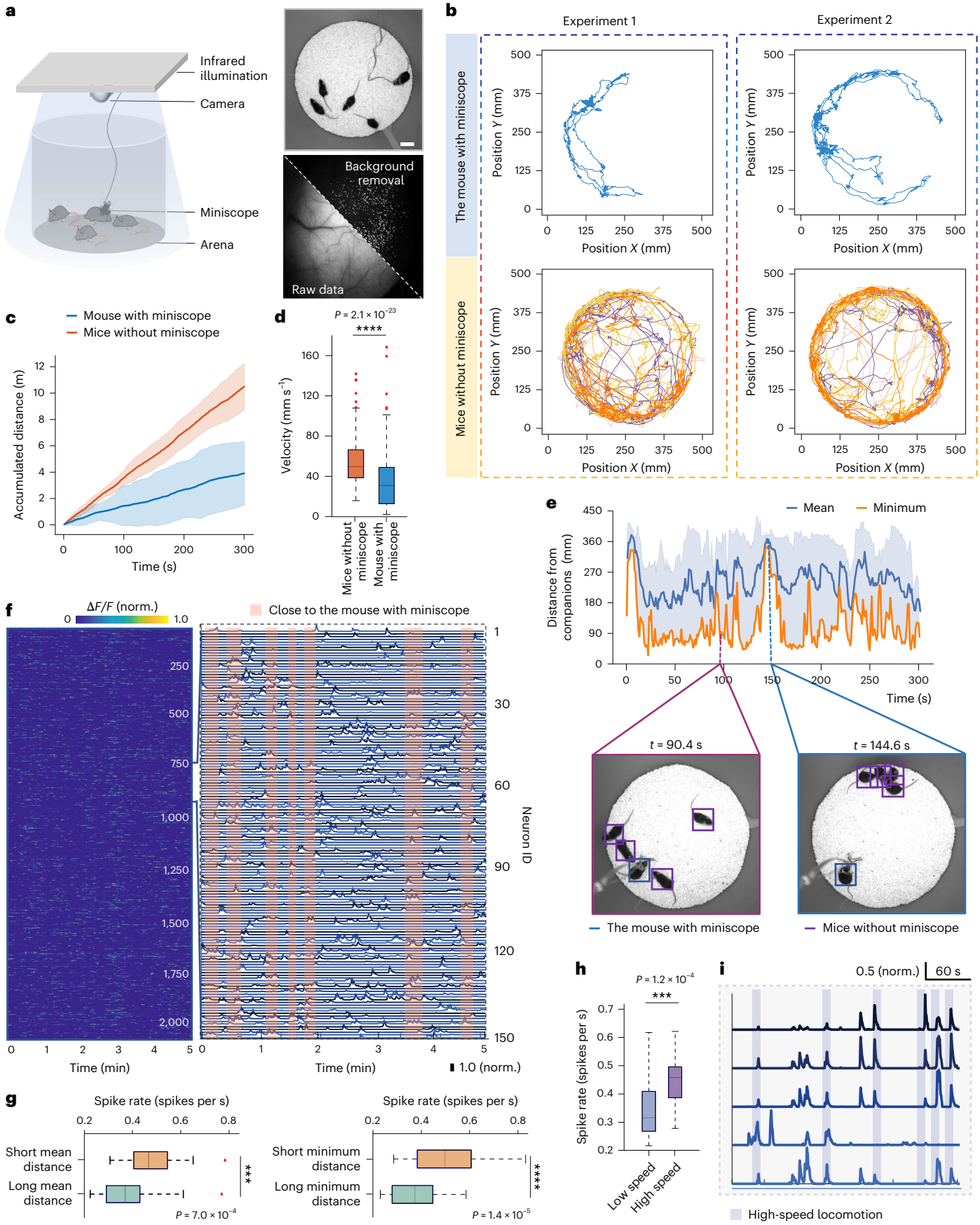
So far, we have demonstrated the performance of our method in rodents. Although rodents are commonly used model animals in laboratories, ethological research is conducted with various animals such as worms, insects and fish[43–45]. To demonstrate the broad applicability of UDMT, we applied it to fruit flies (*Drosophila*), worms (*C. elegans*) and betta fish (*B. splendens*). The body size and behavioral patterns of these species are quite different from those of rodents. For the tracking of flies, we captured the locomotion of multiple flies inside a large culture dish (Fig. 5a,b and Supplementary Fig. 2b). We recorded the movement of 17 flies in the culture dish for 500 s. After extracting their movement trajectories using UDMT, we visualized the trajectories and computed the velocity and acceleration of these flies (Fig. 5c,d and Supplementary Video 6). The results show that the moving velocity and acceleration of these flies decreased after being transferred to the arena, which is probably related to the fact that they finished exploration and gradually became familiar with the new environment. We also performed a more fine-grained analysis by visualizing the trajectories in a three-dimensional (x–y–t) manner. We identified two chasing flies with highly similar trajectories that exhibited a short temporal delay (Fig. 5e).

In *C. elegans*, movement tracking can reveal external perturbations such as mechanical and chemical interventions, as well as internal states such as metabolic health and neural function[46]. We therefore assessed the performance of UDMT in tracking multiple *C. elegans*. Using a stereoscope, we imaged the crawling of seven *C. elegans* inside a culture dish (Fig. 5f and Supplementary Fig. 6a). We tracked the *C. elegans* simultaneously with UDMT and visualized their trajectories and velocities throughout a 22-min recording (Supplementary Video 7). Since both reversal (switching from moving forward to backward) and large-angle turning are important locomotion features of *C. elegans*[47], we derived the frequency of directional change and reversal of these *C. elegans* to quantify how often they change their moving direction (Fig. 5h and Supplementary Fig. 7). Our ethological analysis indicates that these *C. elegans* behaved stably during the recording period and their directional change frequency did not show large fluctuations. Moreover, to compare UDMT with other methods for tracking flies and *C. elegans*, we conducted independent repeated tracking experiments on different numbers of flies and *C. elegans* (Fig. 5i and Extended Data Fig. 6). Comprehensive metrics, including HOTA, MOTA, IDF1 and IDSW, have demonstrated that UDMT can track flies and *C. elegans* with higher accuracy than other methods in different numbers of animals.

Lastly, we used UDMT for the ethological investigation of betta fish, a species that exhibits aggressive behaviors[48]. When two male adult betta fish are close to each other, they will manifest agonistic display and may chase and fight each other. Their aggressive behaviors are commonly adopted to establish dominance and defend territory, particularly during mating season[49]. We recorded the interactions of two betta fish inside an arena (Supplementary Fig. 6b and Supplementary Video 8). The movement trajectories of the two betta fish during the 5-min recording were extracted using UDMT (Extended Data Fig. 7a). By visualizing the trajectories in a three-dimensional coordinate, we can find a segment of trajectories related to aggressive behaviors (Extended Data Fig. 7b). We also calculated the swimming velocity of the betta fish and correlated their aggression behavior with velocity (Extended Data Fig. 7c). Combined with corresponding images, we recognized some fixed patterns of the aggressive behavior of betta fish (Extended Data Fig. 7d). The confrontation between two fish is accompanied by a drop in swimming speed to almost zero, while chasing behavior results in a rapid rise in swimming speed to seven times their average speed. During chasing, the speed of the subordinate fish is much higher than that of the dominant fish, which means that the subordinate fish will

Fig. 4 | Neuroethological analysis of multiple mice combined with head-mounted miniaturized microscopy. a, Neuroethological platform for simultaneous mouse behavior recording and neural calcium imaging (left). Example image of behavior recording (top right; scale bar: 50 mm) and calcium imaging (bottom right; scale bar: 500 μm, maximum intensity projection). Fluorescence background in raw calcium imaging data was removed by a processing pipeline. **b**, Trajectories of the mouse with head-mounted miniaturized microscope (miniscope) and the other four mice without miniscope in two independent experiments. Different colors represent different mice. **c**, Accumulated distance over 5 min of moving. The orange line indicates the mean moving distance of the four free mice and the blue line indicates the moving distance of the mouse with miniscope. Shaded regions denote standard deviation. Quantitative analysis was performed on five independent video recordings (64-Hz frame rate, 19,380 frames per video, $n = 5$). **d**, Tukey box-and-whisker plots showing 1-s averaged speed of the mouse carrying the miniscope and the mean speed of the four free mice. Box plots were constructed from $n = 302$ time bins. Statistical significance was assessed using a one-sided paired *t*-test across corresponding time bins ($P = 2.1 \times 10^{-23}$). **e**, Line plot showing the distance of the mouse with miniscope from the other four mice as a function of time (top) and two representative video frames (bottom). Blue line and shaded region denote mean and range of values, respectively. **f**, Neural activity of 2,131 detected neurons in a 5-min recording. The zoom-in panel shows the calcium traces of 200 neurons. Red shaded areas represent time windows when the distance between the four free mice and the mouse with the miniscope falls below the lower quartile of their average distances. **g**, Tukey box-and-whisker plots showing neuronal spike rate at different mean and minimum spatial distances. Box plots were constructed from $n = 37$ time bins. *P* values were calculated using a one-sided paired *t*-test (left, $P = 7.0 \times 10^{-4}$; right, $P = 1.4 \times 10^{-5}$). **h**, Tukey box-and-whisker plots showing neuronal spike rate at different moving speeds. Box plots were constructed from $n = 37$ time bins. *P* values were calculated using a one-sided paired *t*-test ($P = 1.2 \times 10^{-4}$). **i**, Tuning analysis on a single-neuron level. Calcium traces of neurons that were upregulated by high-speed locomotion. Blue shadows indicate the time of high-speed locomotion. Box plots in **d**, **g**, **h:** the center line indicates the median; box limits indicate the 25th and 75th percentiles; whiskers extend to 1.5 × interquartile range; points indicate outliers. ***$P < 0.001$, ****$P < 0.0001$.

https://doi.org/10.1038/s41592-026-03051-8



https://doi.org/10.1038/s41592-026-03051-8

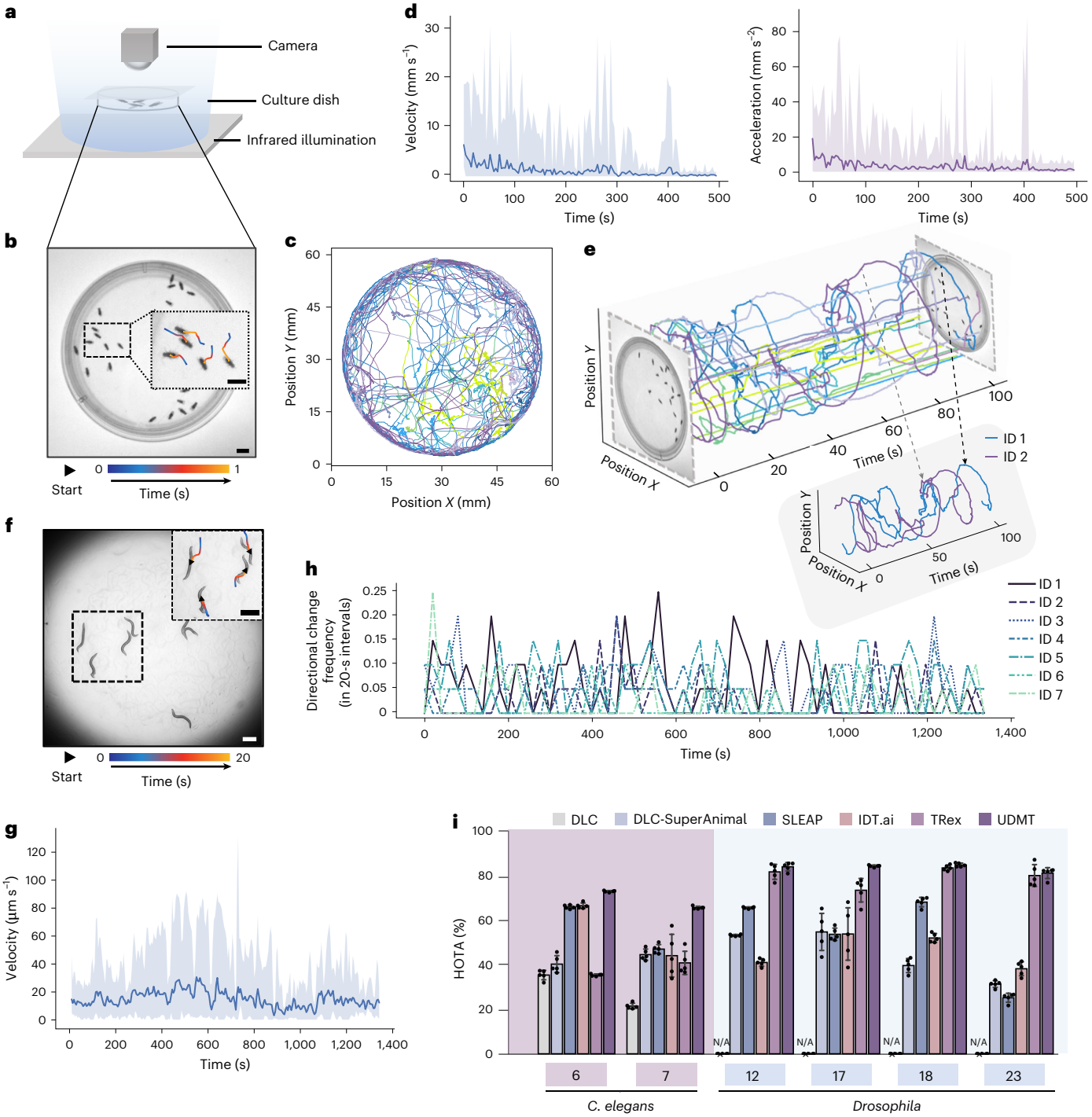


Fig. 5 | General-purpose multi-animal tracking with UDMT. a, Schematic of the recording system for *Drosophila*. **b**, Example image and magnified views of the 17-*Drosophila* dataset. Scale bar: 5 mm. **c**, Projected trajectories of the 17 *Drosophila* during the entire recording period. **d**, Velocity (left) and acceleration (right) of all 17 *Drosophila* averaged over 0.2-s intervals as a function of time. Lines and shaded region denote mean and range of values, respectively. **e**, Three-dimensional (X–Y–t) trajectories of the 17 *Drosophila* in 100 s. The trajectories of two chasing *Drosophila* are shown separately in the inset. The frame rate of the 17-*Drosophila* dataset is 54 Hz and the total number of frames is 26,900. **f**, Example image of the 7-*C. elegans* dataset. Scale bar: 300 μm. **g**, Velocity of all of 7 *C. elegans* averaged over 10-s intervals as a function of time. Lines and shaded region denote mean and range of values, respectively. The frame rate of the 7-*C. elegans* dataset is 10 Hz and the total number of frames is 13,550 frames. **h**, Directional change frequency of 7 *C. elegans*. The directional change frequency was averaged over 20-s intervals. **i**, Comparing the performance of UDMT with other tracking methods on *C. elegans* and *Drosophila* datasets. The number of animals is indicated at the bottom of each bar plot. Bars indicate mean HOTA values across recordings, and error bars indicate standard deviation. Black dots represent individual recordings ($n = 5$ video segments per dataset). N/A means that DLC fails to track *Drosophila* datasets.

escape quickly and then the dominant fish will follow at a relatively slow speed. Such a chasing behavior is aimed at expelling enemies, rather than predation, in which the predator must pursue the prey fiercely with a comparable speed.

Discussion

In this study, we present UDMT, an unsupervised deep transfer learning method for accurate and stable multi-animal tracking without requiring training annotations. UDMT is an unsupervised transfer-learning

method, as reflected in two aspects. From a practical perspective, UDMT can be trained or adapted on the target recordings without manual annotations, including labeled locations, identities, trajectories or keypoints. From a methodological perspective, the core mechanism of UDMT is inherently unsupervised, as it is driven by a cycle-consistency objective without labels. Segmentation foundation models are used only during inference to provide auxiliary cues and are not fine-tuned, and UDMT is initialized from a pretrained model rather than random weights to improve stability and performance during unsupervised training. Overall, UDMT achieves state-of-the-art performance and can track multiple animals accurately under various challenging conditions, including similar appearances, frequent interactions, crowding and occlusion.

Although we tested tracking on up to 23 animals, this is not the upper limit. Since the performance of UDMT is not related only to the number of animals, but also depends on the image resolution and frame rate, improving image resolution and frame rate could allow our method to track more animals. It is noteworthy that the generalization ability of our method is constrained when applied to different species or varying experimental settings. To obtain optimal tracking performance, we recommend training customized models for specific species and experiment conditions.

There are still several avenues to explore in the future. While our current method can simultaneously track the locomotion of multiple animals, it cannot track multiple keypoints on the animal. As the amount of annotation required for keypoint tracking is several times higher than that of position tracking, developing unsupervised methods for keypoint tracking is important for quantitative ethology[17]. Such an advancement would promote fine-grained studies related to animal posture. Moreover, unsupervised learning can eliminate the reliance on task-specific annotations, thus allowing deep learning models to learn from the vast amount of unlabeled data. Inspired by recent progress in foundation models[33,50], we envision designing comprehensive foundation models pretrained on large-scale datasets to realize generalized multi-animal keypoint tracking and behavior profiling, which could provide a versatile, scalable and high-performance solution to address the limitations of current task-specific methods. In terms of applications, we anticipate extending our method to the field of ecology to study the behavior of animals in the wild. Ecological systems exhibit more complex environments than laboratory settings, including streams, coral reefs or forests, and are characterized by biodiversity and large-scale populations[51]. Applying our method to ecological studies such as tracking fish and bird flocks may require appropriate refinements. At the microscale, our unsupervised multi-object tracking method may play an important role in tracking cells, organelles and particles. However, these microscopic objects have different properties from macroscopic objects. For example, division and apoptosis can occur during cell movement. Incorporating more specific rules and constraints is expected to improve its applicability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03051-8.

References

1. Marques, J. C., Li, M., Schaak, D., Robson, D. N. & Li, J. M. Internal state dynamics shape brainwide activity and foraging behaviour. *Nature* **577**, 239–243 (2019).
2. Anderson, D. J. Circuit modules linking internal states and social behaviour in flies and mice. *Nat. Rev. Neurosci.* **17**, 692–704 (2016).
3. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. *Nat. Neurosci.* **23**, 1537–1549 (2020).
4. Mathis, M. W. & Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **60**, 1–11 (2020).
5. Marchant-Forde, J. N. The science of animal behavior and welfare: challenges, opportunities, and global perspective. *Front. Vet. Sci.* **2**, 16 (2015).
6. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
7. Graving, J. M. et al. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* **8**, e47994 (2019).
8. Han, Y. et al. Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nat. Mach. Intell.* **6**, 48–61 (2024).
9. Biderman, D. et al. Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling and cloud-native open-source tools. *Nat. Methods* **21**, 1316–1328 (2024).
10. Pereira, T. D. et al. SLEAP: a deep learning system for multi-animal pose tracking. *Nat. Methods* **19**, 486–495 (2022).
11. Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. H. & de Polavieja, G. G. idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nat. Methods* **16**, 179–182 (2019).
12. Walter, T. & Couzin, I. D. TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife* **10**, 64000 (2021).
13. Marks, M. et al. Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nat. Mach. Intell.* **4**, 331–340 (2022).
14. Lauer, J. et al. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* **19**, 496–504 (2022).
15. Vogg, R. et al. Computer vision for primate behavior analysis in the wild. *Nat. Methods* **22**, 1154–1166 (2025).
16. Li, X. et al. Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising. *Nat. Methods* **18**, 1395–1400 (2021).
17. Sun, J. J. et al. Self-supervised keypoint discovery in behavioral videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2171–2180 (IEEE, 2022).
18. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 16000–16009 (IEEE, 2022).
19. An, X. et al. Unicom: universal and compact representation learning for image retrieval. In *International Conference on Learning Representations* (2023).
20. Li, X., Zhang, Y., Wu, J. & Dai, Q. Challenges and opportunities in bioimage analysis. *Nat. Methods* **20**, 958–961 (2023).
21. Li, X. et al. Unsupervised content-preserving transformation for optical microscopy. *Light Sci. Appl.* **10**, 44 (2021).
22. Xiang, J. et al. A vision-language foundation model for precision oncology. *Nature* **638**, 769–778 (2025).
23. Chen, Y. & Joo, J. Understanding and mitigating annotation bias in facial expression recognition. In *Proc. IEEE/CVF International Conference on Computer Vision* 14980–14991 (IEEE, 2021).
24. Kirillov, A. et al. Segment anything. In *Proc. IEEE/CVF International Conference on Computer Vision* 4015–4026 (IEEE, 2023).
25. Zhao, T. et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**, 166–176 (2025).
26. Wang, N., Song, Y., Ma, C., Zhou, W. & Liu, W. Unsupervised deep tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1308–1317 (IEEE, 2019).

27. Wang, N., Zhou, W., Wang, J. & Li, H. Transformer meets tracker: exploiting temporal context for robust visual tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1571–1580 (IEEE, 2021).
28. Zou, X. et al. Segment everything everywhere all at once. In *Proc. 37th Int. Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 19769–19782 (Neural Information Processing Systems Foundation, Inc., 2023).
29. Cheng, H. K. & Schwing, A. G. Xmem: long-term video object segmentation with an Atkinson–Shiffrin memory model. In *Proc. European Conference of Computer Vision* (eds Avidan, S. et al.) 640–658 (Springer, 2022).
30. Luiten, J. et al. HOTA: a higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **129**, 548–578 (2020).
31. Bernardin, K. & Stiefelhagen, R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP J. Image Video Process.* **2008**, 1–10 (2008).
32. Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. European Conference of Computer Vision* (eds Hua, G. et al.) 17–35 (Springer, 2016).
33. Ye, S. et al. SuperAnimal pretrained pose estimation models for behavioral analysis. *Nat. Commun.* **15**, 5165 (2024).
34. Li, X. et al. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nat. Comput. Sci.* **3**, 1067–1080 (2023).
35. Zhang, G. et al. Bio-friendly long-term subcellular dynamic recording by self-supervised image enhancement microscopy. *Nat. Methods* **20**, 1957–1970 (2023).
36. Li, X. et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nat. Biotechnol.* **41**, 282–292 (2022).
37. Liberti, W. A. III, Schmid, T. A., Forli, A., Snyder, M. & Yartsev, M. M. A stable hippocampal code in freely flying bats. *Nature* **604**, 98–103 (2022).
38. Zong, W. et al. Large-scale two-photon calcium imaging in freely moving mice. *Cell* **185**, 1240–1256 (2022).
39. Wallace, D. J. & Kerr, J. N. D. Circuit interrogation in freely moving animals. *Nat. Methods* **16**, 9–11 (2019).
40. Zhang, Y. et al. A miniaturized mesoscope for the large-scale single-neuron-resolved imaging of neuronal activity in freely behaving mice. *Nat. Biomed. Eng.* **8**, 754–774 (2024).
41. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
42. Li, A. et al. Twist-free ultralight two-photon fiberscope enabling neuroimaging on freely rotating/walking mice. *Optica* **8**, 870–879 (2021).
43. Atanas, A. A. et al. Brain-wide representations of behavior spanning multiple timescales and states in *C. elegans*. *Cell* **186**, 4134–4151 (2023).
44. Lesser, E. et al. Synaptic architecture of leg and wing premotor control networks in *Drosophila*. *Nature* **631**, 369–377 (2024).
45. Huang, K. H. et al. A virtual reality system to analyze neural activity and behavior in adult zebrafish. *Nat. Methods* **17**, 343–351 (2020).
46. Gray, J. & Lissmann, H. W. The locomotion of nematodes. *J. Exp. Biol.* **41**, 135–154 (1964).
47. Hardaker, L. A., Singer, E., Kerr, R., Zhou, G. & Schafer, W. R. Serotonin modulates locomotory behavior and coordinates egg-laying and movement in *Caenorhabditis elegans*. *J. Neurobiol.* **49**, 303–313 (2001).
48. Kwon, Y. M. et al. Genomic consequences of domestication of the Siamese fighting fish. *Sci. Adv.* **8**, eabm4950 (2022).
49. Oldfield, R. G. & Murphy, E. K. Life in a fishbowl: space and environmental enrichment affect behaviour of *Betta splendens*. *Anim. Welf.* **33**, e1 (2024).
50. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
51. Sayin, S. et al. The behavioral mechanisms governing collective motion in swarming locusts. *Science* **387**, 995–1000 (2025).

Methods

Behavioral recording and data annotation

To evaluate how UDMT performs across species, imaging conditions, experimental conditions and other properties of behavioral recordings that may affect tracking performance, we built a collection of diverse animal datasets of behavioral recording (Supplementary Table 2). All experiments involving animals were performed in accordance with the institutional guidelines for animal welfare and have been approved by the Animal Care and Use Committee of Tsinghua University.

Annotation workflow. Human annotators were instructed to label the downsampled original video, assigning an ID to each animal and marking the minimal bounding box around its body parts. We approximated the center of the minimal bounding box as the centroid of the animal, representing its position.

Mouse. The black-mouse dataset was used to assess tracking performance at different numbers of animals, recording frame rate and recording durations. The white-mouse dataset was used to evaluate tracking performance under low-contrast imaging conditions where animal brightness and color are very close to the background. Male C57BL/6J mice were housed in a temperature- and humidity-controlled environment on a 12-h reversed light–dark cycle, with unrestricted access to food and water. Groups of five mice were housed per cage and used for experiments at 2.5 to 4 months of age. Behavioral assays were conducted during the active (dark) phase of mice, from 12:00 to 17:00. To ensure environmental standardization and minimize intra-cage disturbance, mice were individually housed 24 h before behavioral testing. Age-matched animals were used across experimental groups to eliminate potential age-related confounders.

During behavioral recording, mice moved freely in an open-field arena with a diameter of 50 cm and a height of 30 cm. For experiments in complex environments, we placed a water dispenser, a food source, a small fence and a wooden stump for shelter inside the original arena. Videos were captured under infrared illumination (35 × 35 cm^2) using a Mindvision camera (cat. no. MV-SUF401GM) at 40–94 Hz. This camera provides a maximum spatial resolution of 2,048 × 2,048 pixels and a frame rate of 88 Hz at full resolution. The recording frame rate can be increased by decreasing the image resolution using the control software. Infrared brightness was regulated by an analog controller (cat. no. JH-AP60-2C). The custom-assembled acquisition workstation was equipped with an Intel Xeon CPU, 32 GB of RAM, a Seagate BarraCuda 1-TB hard disk drive for data storage and an NVIDIA GeForce GTX 1070Ti graphics processing unit (GPU, 8 GB of memory). The raising and recording methods of rats are the same as those used for mice. For each video, human annotators labeled one frame every 300 frames, obtaining 712 labeled frames on the mice and rat datasets.

Mouse with head-mounted miniaturized microscope. For neural functional imaging, we used transgenic mice bred by crossing Ras-grf2-2A-dCre with Ai148 (TIT2L-GC6f-ICL-tTA2)-D strains, expressing Cre-dependent GCaMP6f genetically encoded calcium indicator. Craniotomy surgeries were performed as previously described[40]. Briefly, mice were anesthetized with 1.5% isoflurane (v/v in oxygen), and a 4.0-mm craniotomy was created using a skull drill. After removing the skull piece, a coverslip was placed over the exposed area, and a titanium headpost was affixed to the skull to stabilize the miniaturized microscope.

The miniaturized microscope, optimized for mesoscale imaging, offers an FOV of 3.6 × 3.6 mm^2, a lateral resolution of 4 µm and a depth of field of 300 µm, with a total weight of 2.5 g. Two weeks after surgery, mice were re-anesthetized and positioned in a stereotactic frame for baseplate attachment. The baseplate was manually aligned for optimal FOV centering on the sensor, then secured with dental cement and Krazy Glue. After the glue was cured, the miniaturized microscope was detached, and the mice were returned to their home cages, now ready for imaging experiments. Before each behavioral experiment, mice were lightly anesthetized (0.5–1% isoflurane in oxygen) and head-fixed to clean the cranial window and mount the miniaturized microscope. A blue light-emitting diode was then activated for fluorescence excitation. Focusing was achieved by adjusting the top housing relative to the bottom housing until neurons were clear, after which the setup was locked. Once focused, mice were released into the open-field arena. Recordings were captured at 4 Hz with a resolution of 2,592 × 1,944 pixels.

Drosophila. *Drosophila* (wild-type Canton S; sex not determined) were grown on standard cornmeal media (BuzzGro from Scientiis) in 25 °C incubators with a 12/12-h light–dark cycle. The behavioral experiments setup was placed in a dedicated experimental room with controlled humidity (60%) and temperature (25 °C). The *Drosophila* were placed in a covered polystyrene petri dish for behavioral recording, which had a diameter of 54 mm and a height of 2 mm. Infrared illumination was positioned directly above the culture dish to ensure consistent lighting within the central region (Supplementary Fig. 2b). A white plastic sheet was placed below the arena to increase the contrast between *Drosophila* and the background. Videos were recorded with the same camera used for mice experiments and obtained with standard top-view recording. Black shade cloths were positioned around the camera to minimize reflections on the glass covering the arena. To transfer *Drosophila* into the arena, they were briefly anesthetized on ice before being gently placed into the setup. Recording began after they regained consciousness and normal activity. Human annotators labeled one frame every 150 frames per video, resulting in 629 labeled frames for the *Drosophila* datasets.

C. elegans. *C. elegans* were cultured on nematode growth medium plates seeded with OP50 *Escherichia coli* cells. The wild-type Bristol N2 strain was provided by the Caenorhabditis Genetics Center and maintained at 20 °C. Unless otherwise stated, hermaphrodites were used for all experiments. To pick *C. elegans* using a worm picker (platinum wire), we slowly lowered the tip of the wire and gently swiped it along the side of the worm, then transferred the worm to a 60-mm nematode growth medium plate seeded with an OP50 lawn. For *C. elegans* imaging, a Nikon Ti2-E inverted stereoscope was utilized (Supplementary Fig. 6a). The system was equipped with a ×2 objective lens (0.1 numerical aperture, 8.5-mm working distance), providing high-resolution images suitable for detailed behavioral analysis. For each video of *C. elegans*, human annotators labeled one frame every 150 frames, resulting in 196 labeled frames.

B. splendens. Male *B. splendens* used in this experiment were raised in 600-ml glass flasks at 26 ± 2 °C under a 12/12-h light–dark cycle. Each fish was placed in a separate tank within a circulating aquarium system. Fish were fed twice daily and their average length was about 5.2 cm. For behavioral recordings, two male betta fish were transferred to a 20 × 30 × 30-cm^3 transparent glass tank filled with water with a depth of 2.5 cm (Supplementary Fig. 6b). Videos were recorded with the same camera as in the mice setup. To prevent reflections on the water surface, infrared illumination was positioned beneath the fish tank, which was supported by a transparent acrylic plate to maintain a distance of approximately 30 cm from the light source. Additionally, a thick layer of frosted cellophane was placed at the bottom of the tank to ensure uniform illumination.

Network architecture

The network architecture of UDMT retains the topology established in TrDiMP[27], including a feature extraction module, a transformer encoder module, a transformer decoder module and a discriminative filter. We employed ResNet-50[52] as the backbone network within the

feature extraction module to derive the initial feature maps for both the template and search regions. The primary operation within the transformer encoder is self-attention, designed to enhance the features derived from a dynamically updated template ensemble. Notably, the self-attention blocks in both the encoder and decoder share weights, enabling the transformation of template and search embeddings into a unified feature space, thus facilitating subsequent cross-attention computations. We computed the cross-attention based on the search feature and template feature. The enhanced search features extracted by the spatial attention mechanism can better highlight potential target areas. Ultimately, we constructed a discriminative filter from template features and convolved it with search features to accurately localize the target.

Training and inference

The training process starts with semi-automatic segmentation of all target animals in the first frame using a fine-tuned foundation model[28], allowing for manual refinement via the GUI if required. A small amount of manual intervention may be required to correct initialization biases. Then, a pretrained video object segmentation model is used for animal segmentation throughout the entire video[29]. The masks are utilized in the localization refining module to improve tracking accuracy. During the forward tracking phase of training, we employed the raw target positions from the preceding ($t-1$), current (t) and subsequent ($t+1$) frames to crop object regions, which serve as the templates. t is randomly sampled from all frames. Frames are cropped centered on the target position and the side length is determined by the search region size. If the cropped region extends beyond the image border, it is padded with the average intensity of the entire image, which closely approximates the background intensity. The cropped regions from $t+\alpha-1$, $t+\alpha$ and $t+\alpha+1$ frames serve as the search frames, where α is an integer randomly sampled from the range of 1 to 10. Training image patches were extracted by sampling random translations relative to the target annotations for data augmentation. The templates and search frames will be simultaneously fed into the network to locate the position of the target animal within the search frames. In the backward tracking phase, the search frames from the previous step are used as templates, and processed by the same network. The value of the loss function is computed using the initial positions and the corresponding positions of backward tracking in $t-1$, t and $t+1$ frames. Simply using the difference between ground-truth target scores and target confidence scores as the loss function biases the learning process toward negative data samples, rather than enabling the model to achieve optimal discrimination. Moreover, using naive differences cannot address the problem of data imbalance between targets and backgrounds. To address this issue, we modified the hinge loss[53] based on the principle of support vector machines to train UDMT models in an end-to-end manner. The loss function for each label and prediction score pair is formulated by

$$I(s,z)=\begin{cases} s-z, & z>\mathrm{T} \\ \max(0,s), & z\le \mathrm{T} \end{cases}$$

In this context, the threshold T represents the target and background regions based on the label confidence value z. For the target region where $z>\mathrm{T}$, we computed the difference between the predicted confidence score s and the label z. In contrast, for the background region where $z\le\mathrm{T}$, we penalized only positive confidence values. T was set to 0.05 in our experiment.

During the inference phase, given an annotated initial frame, subsequent frames are processed sequentially through the network. For preprocessing, each input frame is normalized by subtracting its minimum intensity value and then dividing by the difference between the maximum value and the minimum to handle the intensity variation across different videos. To leverage temporal information and adapt to changes in target appearance optimally, the template ensemble in the transformer module will be updated dynamically. Specifically, for every five frames, the earliest template in the ensemble is discarded, and the current template feature is appended. The feature ensemble is maintained at a maximum of 15 templates. To process the first frame, we employed data augmentation strategies (including blur, rotation and shift) to construct an initial template ensemble containing 15 samples. Upon updating the template ensemble, we computed the new encoded template feature using our transformer encoder. While the transformer encoder is sparsely utilized (that is, every five frames), the transformer decoder is engaged for each frame, generating per-frame search features by propagating representations and attention cues from preceding templates to the current search patch.

All models were trained using the Adam optimizer[54], with an exponential decay rate of 0.9 for the first moment and 0.999 for the second moment. The learning rates for the feature extraction module, transformer module, filter initialization and filter optimization module were 0.00005, 0.001, 0.00005 and 0.0005, respectively. The ST-Net was initialized using weights provided in TrDiMP, whereas the discriminative filter was trained from scratch. The search area factor was set to 3.0, and the default number of samples (image pairs) per epoch was set to 5,000. We used GPUs to accelerate the training and testing process. The batch size for all experiments was 40, which required 16.93 GB of memory per GPU when training in parallel on three GPUs. Smaller batch sizes can reduce memory demands. Training the network with a video comprising approximately 4,000 frames is sufficient to ensure satisfactory performance (Supplementary Fig. 8a). Generally, the optimal tracking performance is achieved at the 10th epoch and keeps stable within 20 epochs (Supplementary Fig. 8b). Training the model for 20 epochs on a typical dataset (with a batch size of 16) using a single Nvidia GeForce RTX 3090 GPU (24 GB of memory) takes approximately 2 h. On an already trained network, processing 16,000 frames (652 × 636 pixels, containing five animals) requires about 1,800 s. Training time can be further reduced by utilizing more powerful GPUs or parallelizing computations on multiple GPUs. PyTorch was used to construct the network and implement all operations. The tracking result of each video was saved as a separate txt file.

Semi-automatic initialization

Reliable multi-animal tracking requires precise annotation of the initial location for each individual. To this end, we developed a semi-automatic initialization approach that combines model-driven instance segmentation with optional manual refinement. In the first frame of the input video, animal instances are segmented by a domain-adapted foundation model[28] fine-tuned on a curated dataset encompassing six organism types (black mouse, rat, white mouse, *C. elegans*, *Drosophila* and fish). Fine-tuning was performed for 50 epochs using the Adam optimizer (learning rate 1×10^{-4}, weight decay 0.05), while the backbone, language encoder and pixel decoder were frozen to preserve pretrained representations. During inference, animal locations are extracted from the first frame using prompt-free panoptic segmentation. Users may supplement missed detections to ensure complete initialization. All steps are fully integrated within our GUI.

Localization refining module

The localization refining module was designed to improve the accuracy of animal localization in real-time. Without the refining module, when a position deviation occurs during tracking, the error will accumulate frame-by-frame and degrade the accuracy of long-term tracking. Such inaccurate or unstable outputs have been shown to adversely affect downstream analyses[55]. For localization refining, all video frames and the initial positions of animals are sent into a pretrained model to segment the masks of all animals. To avoid the bias of network prediction, the refining module uses the center of gravity of corresponding segment masks to replace the original localization of the network.

The refined positions will be used to update subsequent tracking. The refining process is applied only for those segmentation masks whose areas are between 50% and 120% of the initial animal area, ensuring that the mask corresponds to a single animal. The initial animal area is obtained by averaging the area of all animals in the first frame. When multiple animals are in contact, the area of each animal is estimated by dividing their total area by the number of animals within that region. For some concave-shaped animals such as *C. elegans*, their centroid may not be located within segmentation masks. In these cases, the centroid of the smallest outer rectangle of each segmentation mask is used as the animal position.

ID correction module

To reduce cumulative ID errors, we developed an ID correction module leveraging bidirectional tracking. The detailed workflow of ID correction is shown in Extended Data Fig. 3. First, if the trajectory difference index[56] between any two animals in a frame falls below an empirical threshold (0.8 in our work) and the distance between these two animals is less than the initial animal size, it indicates that this frame (lost target frame) contains a lost target. Next, UDMT detects an abrupt jump in movement speed (based on variance and max/mean) and a sudden drop in localization confidence to determine which animal is assigned the incorrect ID through majority voting. Finally, the position of the lost animal is re-localized by detecting the animal mask (segmented by the localization refining module in the lost target frame) without ID. The trajectory of the lost target is fixed through backward tracking. Besides, if the position of a specific ID stays outside the animal mask (off-target localization) for more than 3 s, it is identified as a missing target. The same procedure will be applied to correct it.

Automatic parameter tuning module

Tracking accuracy is highly related to search region size, which is codetermined by the search region scale and target size bias (Supplementary Table 1a). The target size is the side length of the tracking box and is defined as the initial animal size plus the target size bias. The search region scale is the scale factor of the search region size relative to the target size. The search region size is equal to the target size multiplied by the search region scale. The search region scale (1.5, 2 or 2.5) and target size bias (−10%, 0% or 10% of the initial target size) are automatically adjusted based on proposed evaluation metrics, including the number of ID corrections, off-target localizations and missing targets. They are strongly correlated to tracking performance (Supplementary Table 1b) and can be calculated without ground truth. The number of ID corrections reflects how often the ID correction module is used and has the strongest correlation with tracking performance. In most cases, fewer ID corrections indicate that the parameters used are more suitable. The number of off-target localizations refers to the cumulative number of instances where the predicted animal position of UDMT falls outside the animal mask segmented by the localization refining module. Off-target localization lasting more than 3 s is considered to be a missing target, necessitating the ID correction module to reassign an ID to it. The number of missing targets can reflect tracking accuracy.

The detailed flowchart of automatic parameter tuning is shown in Extended Data Fig. 2. We adopted conditional iterative optimization to find the best parameters (the target size bias and search region scale) that can lead to the smallest evaluation metrics. At the beginning of the iteration, we initialized the current optimum to a very large value (1,000). After processing each frame, the number of ID corrections is compared with the current optimum. If the value is larger than the current optimum, the tracking using this search region size is terminated. If the number of ID corrections equals the current optimum, further conditions are evaluated sequentially, including the number of off-target localizations and missing targets. The tracking process continues if none of these metrics is larger than the current optimum or the number of ID corrections is smaller than the current optimum. If the final frame is processed, the optimal evaluation metrics are updated. The iteration continues over all optional values of the search region size until finished. Additionally, the processing time serves as a constraint on algorithm efficiency. If other parameters have the same values, the parameter set with the shortest processing time will be adopted. Generally, a video segment of about 1 min is sufficient for automatic parameter tuning. The best parameters will be used in both training and inference.

Software

The results reported in this paper were produced using UDMT (version 1.1.1). The software environment includes Python 3.8, PyTorch 1.7.1 and CUDA 11.8. A step-by-step tutorial is available at https://cabooster.github.io/UDMT/Tutorial/. All relevant pretrained models are listed in the tutorial page. The core dependencies include TorchVision (version 0.8.2), Torchaudio (version 0.7.2), NumPy (version 1.24.4), SciPy (version 1.10.1), scikit-learn (version 1.3.2), scikit-image (version 0.21.0), OpenCV-Python (version 4.11.0.86), pandas (version 2.0.3), Matplotlib (version 3.7.5), ImageIO (version 2.35.1), imgaug (version 0.4.0), statsmodels (version 0.14.1), segment-anything (version 1.0), Tensorpack (version 0.11), tqdm (version 4.67.1), requests (version 2.32.3), Pillow (version 10.4.0) and similaritymeasures (version 1.2.0). The GUI was built using PySide6 (version 6.3.1).

Ablation study

To evaluate the effectiveness of the transformer architecture and the three modules in UDMT (localization refining, ID correction and automatic parameter tuning), we conducted ablation studies using behavioral recordings of seven mice (67-Hz frame rate, 29,550 frames, $n = 5$ video segments). DiMP[57] was used as the baseline representing CNNs, which has a similar architecture as UDMT and substitutes the transformer blocks with convolutional blocks. The backbone network of DiMP was initialized with the open-source pretrained weights (DiMP-50) publicly available at https://github.com/visionml/pytracking/blob/master/MODEL_ZOO.md. The model was fine-tuned for 20 epochs. For the experiments removing the ID correction module, we randomly reassigned IDs to the missing target of two intersecting trajectories. Backward tracking was retained to restore missing trajectories. For those experiments without automatic parameter tuning, we selected the initial target size and typical search region scale (2.0) as the parameters to obtain representative tracking performance.

Method comparison

We compared the performance of UDMT with five baseline methods: DLC[14], DLC-SuperAnimal[33], SLEAP[10], IDT.ai[11] and TRex[12]. All methods were implemented with publicly available code provided by companion papers. Human annotators were instructed to label keypoints for each dataset to generate the training dataset for supervised methods (DLC, DLC-SuperAnimal and SLEAP). For mice, eight keypoints (snout, left ear, right ear, shoulder, three spine points and tail base) were annotated for 100 frames per video. For *Drosophila*, eight keypoints (head, left eye, right eye, thorax, left midleg, right midleg, left hindleg and right hindleg) were annotated for 30 frames per video. For *C. elegans*, seven keypoints (head, five body points and tail) were annotated for 50 frames per video. Annotated frames were sampled from original videos using uniform clustering. DLC was trained using a ResNet backbone pretrained on ImageNet, with a randomly initialized pose estimation head, following the standard DLC pipeline. DLC-SuperAnimal was initialized from the TopViewMouse SuperAnimal model and subsequently fine-tuned on our dataset to adapt the pretrained keypoint representations to the target animals. For the training of SLEAP, the maximum number of instances was set to match the number of animals. The second spine point, thorax and third body point were designated as the anchor points for mice, *Drosophila* and *C. elegans*, respectively. For

the inference of SLEAP, a flow-shift tracker was applied. The maximum number of instances and tracks were both set to the number of animals to ensure realistic tracking. The 'connect single track breaks' option was applied for all videos. For DLC, DLC-SuperAnimal and SLEAP, the anchor point representing the centroid of each animal was used to localize the animal. For IDT.ai., we applied background subtraction for all videos and manually adjusted intensity and area thresholds using the GUI to segment individual animals. For TRex, we followed the official documentation and used the GUI to assist parameter tuning. Detection and tracking parameters were manually adjusted for each video to optimize performance. The 'Visual Identification' module was applied where appropriate, and the final result for each video was selected based on overall tracking quality. Missing values of these methods were estimated using linear interpolation based on the nearest available data. All hyperparameters not mentioned here were set as default values.

Locomotion analysis

To obtain the instantaneous velocity of a given frame, we computed the Euclidean distance between the animal coordinates of the current frame and the next frame, and then divided by the time interval between the two frames. To calculate acceleration, we calculated the difference in the instantaneous velocity of two adjacent frames and divided by the time interval between the two frames. To identify directional changes, we resampled the trajectory at ten-frame intervals to convert it into a path of contiguous line segments with a constant time gap. Turning angles were defined as the smallest angle between adjacent segments.

Calcium imaging analysis

Raw calcium imaging data from the head-mounted miniaturized microscope were motion-corrected by NormCorre[58]. The data were then processed by DeepWonder[59] to extract neuronal masks and calcium traces. DeepWonder was fine-tuned to make it suitable for miniaturized microscopes, following the guidelines provided at https://github.com/yuanlong-o/Deep_widefield_cal_inferece/tree/main/Alternative/DeepWonder. For spike inference, we used the MLspike algorithm[60], which was ranked first in the Spikefinder challenge[61]. Before being fed into the spike inference pipeline, all calcium traces were divided by their mean values for normalization. The recommended parameters for the GCaMP6f calcium indicator were used to ensure optimal performance in spike inference.

Neuroethological analysis

For the mouse with the miniaturized microscope, high moving speed was defined as any speed larger than the mean speed. Short distance was defined as any distance below the lower quartile of the average distance between the mouse with the miniaturized microscope and its companion, which means that they were close to each other. These events must last longer than 5 s to be recognizable. Neurons exhibiting significantly increased average activity during the high-speed movement were classified as upregulated neurons ($P < 0.05$, two-sided Wilcoxon test). We performed an independent sample t-test to assess whether there was a significant difference in the speed of the mouse with and those without the miniaturized microscope. We also assessed whether the speed of the mice and their distance influenced their neuronal spike rate. These variables were divided into two groups: one consisting of values greater than the upper quartile and the other consisting of values less than the lower quartile. Before the t-test, we verified the homogeneity of variance between the two groups using Levene's test.

To quantify social-distance-related modulation using an ID-resolved analysis of social proximity, we analyzed neuronal activity of the mouse with the miniscope as a function of its proximity to each individually tracked cage-mate. Inter-animal Euclidean distance was computed frame-by-frame from tracked centroids and converted to millimeters using calibration. 'Close' was defined as distance ≤127 mm (approximately 4/3 body lengths) and 'Far' as distance >127 mm (ref. 62). Proximity bouts shorter than 5 s were discarded to ensure behavioral stability. For each cage-mate analyzed independently, we computed per-neuron mean spike rate within Close and Far bouts and treated neurons as paired observations across conditions. Normality of within-neuron differences was assessed; when appropriate we used a paired t-test, and otherwise a Wilcoxon signed-rank test (two-sided). Resulting P values were visualized as bar plots and indicated with asterisks. All analyses were performed by customized MATLAB scripts.

Performance metrics

Tracking performance is assessed with comprehensive metrics including HOTA, MOTA, IDF1 and IDSW. Among them, the primary metric is HOTA, which is computed by the code provided at https://github.com/JonathonLuiten/TrackEval. It can balance the effect of detection and association into a unified metric for explicit comparison[30]. In terms of detection, its accuracy is quantified by the overall detection accuracy (DetA), measuring the consistency between the network prediction and corresponding ground truth using intersection over union (IOU). Since a predicted object may overlap with multiple ground-truth annotations, the Hungarian algorithm is applied for one-to-one matching. Matched pairs are true positives (TPs), while unmatched predicted detections are false positives (FPs) and unmatched ground-truth detections are false negatives (FNs). For a given IOU threshold α, DetA is calculated as:

$$\mathrm{DetA}_\alpha = \frac{|\mathrm{TP}_\alpha|}{|\mathrm{TP}_\alpha| + |\mathrm{FN}_\alpha| + |\mathrm{FP}_\alpha|}.$$

Association accuracy measures the effectiveness of a tracking model in associating detected objects to correct IDs over time. The Hungarian algorithm is used for matching and IOU is used for quantification. The intersection between two tracks is quantified by true positive associations (TPAs). Objects in the predicted track that are unmatched or matched to other ground-truth annotations are false positive associations (FPAs), while unmatched objects in the ground-truth trajectories are false negative associations (FNAs). The overall association accuracy (AssA) is calculated by averaging the association intersection-over-union (Ass-IOU) across all TPs in the entire dataset:

$$\mathrm{AssA}_\alpha = \frac{1}{|\mathrm{TP}_\alpha|} \sum_{c \in \mathrm{TP}_\alpha} \frac{|\mathrm{TPA}_\alpha|}{|\mathrm{TPA}_\alpha| + |\mathrm{FNA}_\alpha| + |\mathrm{FPA}_\alpha|}(c).$$

For each threshold α, HOTA is defined as the geometric mean of the detection and association accuracy, which ensures detection and association are evenly weighted. The final value of HOTA is obtained by integrating over different thresholds:

$$\mathrm{HOTA} = \int_{0<\alpha\leq 1} \mathrm{HOTA}_\alpha \approx \frac{1}{19} \sum_{\substack{\alpha=0.05 \\ \alpha+=0.05}}^{0.95} \mathrm{HOTA}_\alpha = \frac{1}{19} \sum_{\substack{\alpha=0.05 \\ \alpha+=0.05}}^{0.95} \sqrt{\mathrm{DetA}_\alpha \mathrm{AssA}_\alpha}.$$

MOTA is a metric that is more sensitive to detection than association. It contains two types of detection errors (FNs and FPs) and one type of association error (IDSW). MOTA is computed by summing these three errors, dividing by the number of ground-truth objects (gtDet) and then subtracting from 1:

$$\mathrm{MOTA} = 1 - \frac{|\mathrm{FN}| + |\mathrm{FP}| + |\mathrm{IDSW}|}{|\mathrm{gtDet}|}.$$

IDF1 evaluates the consistency of ID associations by balancing precision and recall. It is particularly effective in measuring association

performance and is sensitive to IDSW, making it suitable for complex or long-duration tracking scenarios. However, IDF1 does not directly account for detection errors such as FPs and FNs, and it needs to be complemented with other metrics such as MOTA or HOTA to provide a more comprehensive evaluation of tracking performance[32]. IDF1 is formulated as:

$$\mathrm{IDF1} = \frac{|\mathrm{IDTP}|}{|\mathrm{IDTP}| + 0.5\,|\mathrm{IDFN}| + 0.5\,|\mathrm{IDFP}|}$$

where IDTP, IDFN and IDFP represent the number of ID TPs, FNs and FPs, respectively. When calculating performance metrics, the bounding box of an animal is defined as the maximum outer square centered at the position of it.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We have no restriction on data availability. All behavioral recordings used in this study are available via Zenodo at https://doi.org/10.5281/zenodo.14580255 (ref. 63). The source data of neuroethological research of freely behaving mice, including behavioral recordings and synchronized calcium imaging data, are available via Zenodo at https://doi.org/10.5281/zenodo.14586425 (ref. 64). Source data are provided with this paper.

Code availability

All relevant resources are readily accessible on our GitHub page at https://cabooster.github.io/UDMT/. The source Python code of UDMT can be found at https://github.com/cabooster/UDMT. The code is distributed under the 'ACADEMIC OR NON-PROFIT ORGANIZATION NONCOMMERCIAL RESEARCH USE ONLY' license (see the LICENSE file in the repository). A frozen copy of the code is available via Zenodo at https://doi.org/10.5281/zenodo.18312420 (ref. 65). Our specially designed user-friendly GUI and accompanying tutorial can be found at https://cabooster.github.io/UDMT/Tutorial/.

References

52. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
53. Bartlett, P. L. & Wegkamp, M. H. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.* **9**, 1823–1840 (2008).
54. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (2014).
55. Weinreb, C. et al. Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nat. Methods* **21**, 1329–1339 (2024).
56. Müller, M. *Information Retrieval for Music and Motion* (Springer, 2007).
57. Bhat, G., Danelljan, M., Gool, L. V. & Timofte, R. Learning discriminative model prediction for tracking. In *Proc. IEEE/CVF International Conference on Computer Vision* 6182–6191 (IEEE, 2019).
58. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: an online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).
59. Zhang, Y. et al. Rapid detection of neurons in widefield calcium imaging datasets after training with synthetic data. *Nat. Methods* **20**, 747–754 (2023).
60. Deneux, T. et al. Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nat. Commun.* **7**, 12190 (2016).
61. Berens, P. et al. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.* **14**, e1006157 (2018).
62. Klibaite, U. et al. Mapping the landscape of social behavior. *Cell* **188**, 2249–2266 (2025).
63. Li, Y. UDMT dataset: Behavioral recordings used for unsupervised multi-animal tracking. *Zenodo* https://doi.org/10.5281/zenodo.14580255 (2024).
64. Li, Y. UDMT dataset: Source data of neuroethological research of freely behaving mice. *Zenodo* https://doi.org/10.5281/zenodo.14586425 (2025).
65. Li, X. cabooster/UDMT: UDMT v1.1.1. *Zenodo* https://doi.org/10.5281/zenodo.18312420 (2026).

Acknowledgements

We thank L. Yuan, G. Xiao and J. Xie from the Department of Automation, Tsinghua University, for assistance with calcium imaging using the head-mounted miniaturized microscope and for providing the rodents used in this study. This work was supported by the National Natural Science Foundation of China (grant no. 62088102 to Q.D., grant nos. 62222508 and 62525506 to J.W., grant no. 32571275 to X.L.), Beijing Natural Science Foundation (grant no. Z240011 to J.W.), the Chinese Postdoctoral Foundation (grant no. 2023M741962 to X.L.), the Tsinghua Shuimu Scholar Program (grant no. 2023SM066 to X.L.), the New Cornerstone Science Foundation through the XPLORER PRIZE (to J.W.), the Beijing Key Laboratory of Cognitive Intelligence (to Q.D.) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (grant no. JYB2025XDXM504 to Q.D. and Z. Li).

Author contributions

Q.D., J.W., Z. Li and X.L. supervised this research. Q.D. and X.L. conceived and initiated this project. Y.L. and X.L. designed detailed implementations, built the recording system and performed experiments. Y.L. developed the Python code, completed the GUI and processed relevant data. Q.Z. and J.F. provided *C. elegans* and *Drosophila*, and assisted with relevant experiments. Y.Z., Z. Lu and X.X. provided the head-mounted miniaturized microscope and gave critical support on its imaging and data processing. Y.L. and X.L. analyzed the data, prepared figures and videos, and made the companion webpage. Y.L., X.L., Q.Z., Z. Li and J.W. participated in discussions about the results. All authors participated in the drafting of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

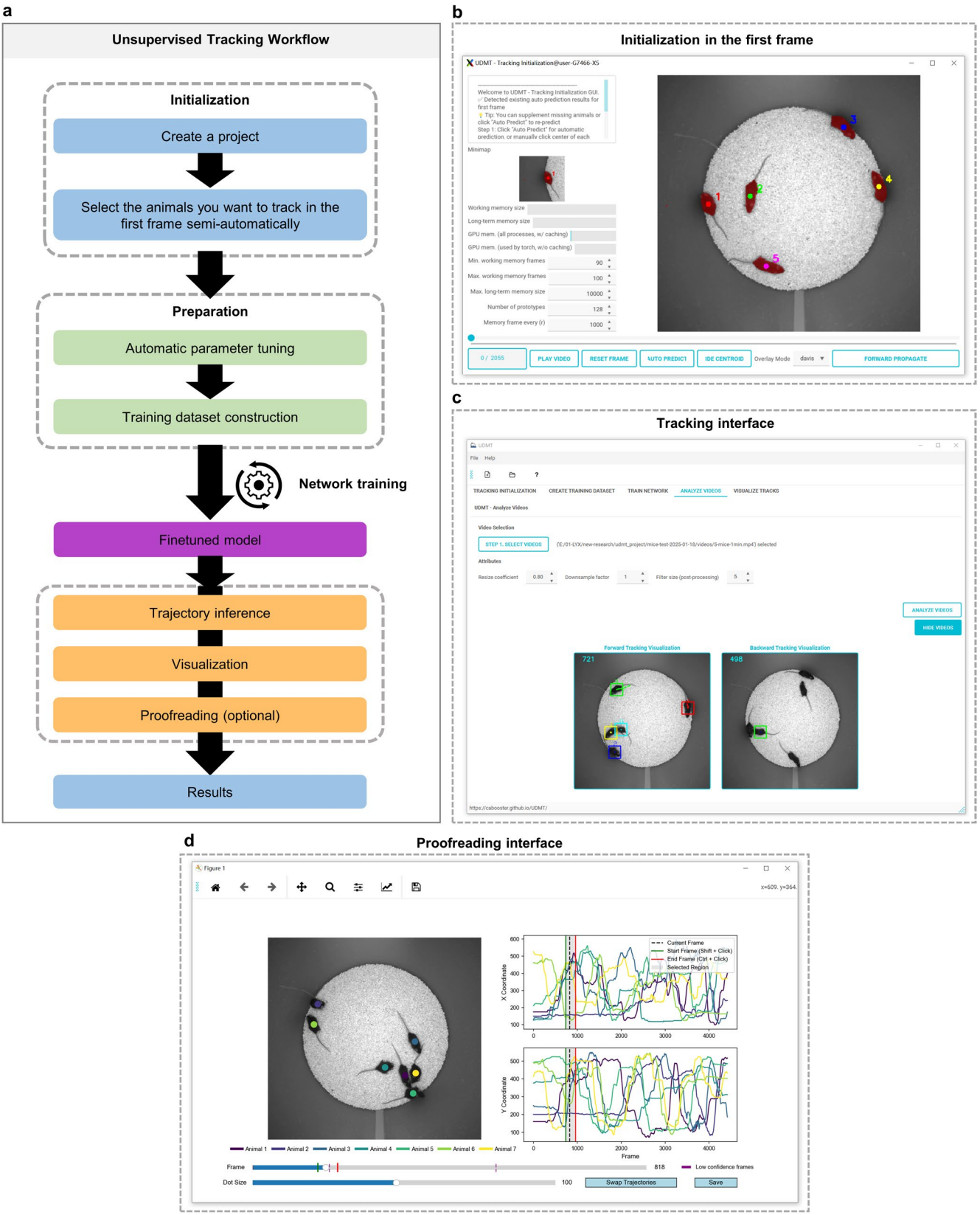
Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03051-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03051-8.

Correspondence and requests for materials should be addressed to Xinyang Li, Ziwei Li, Jiamin Wu or Qionghai Dai.

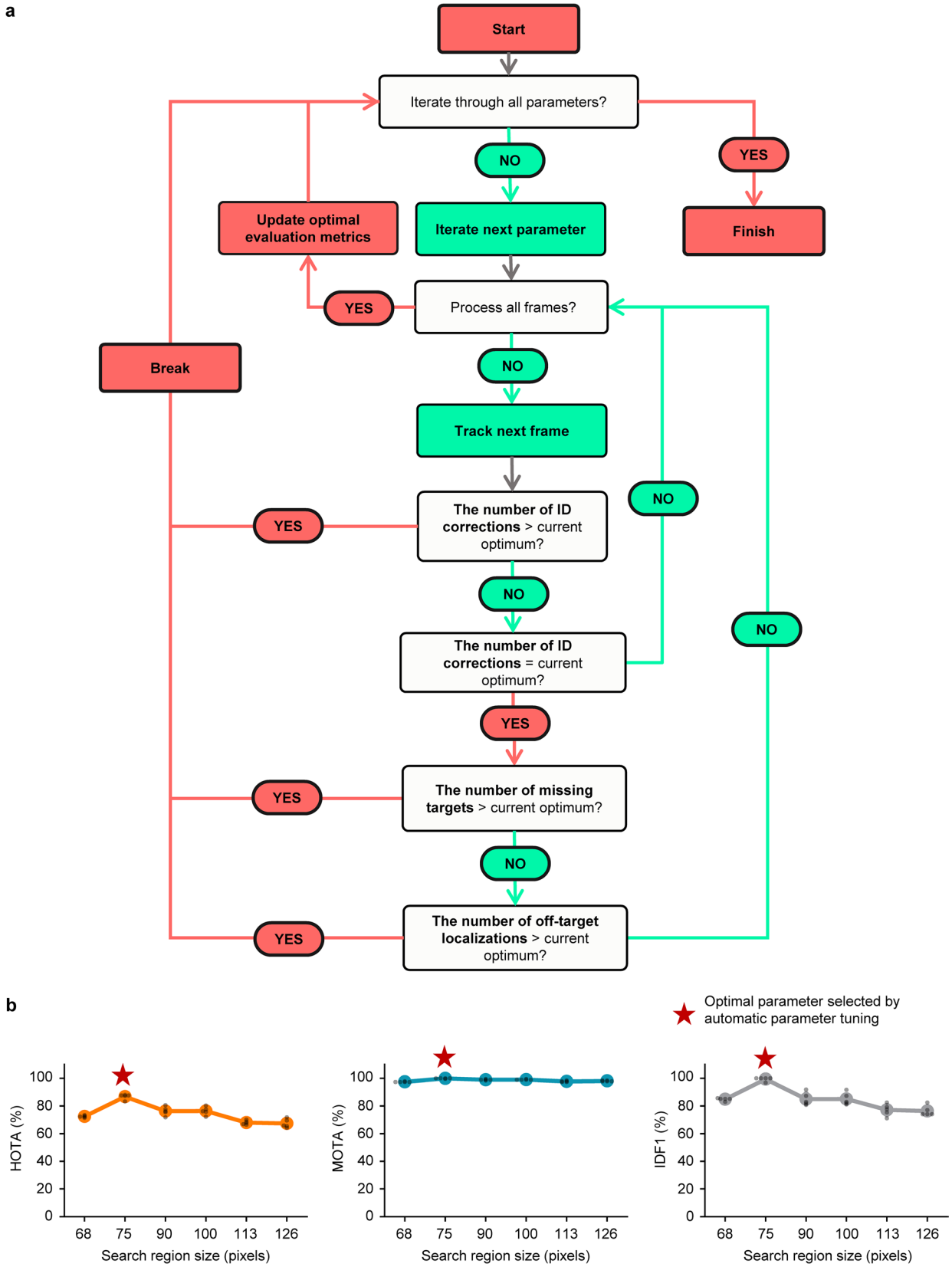
Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-026-03051-8



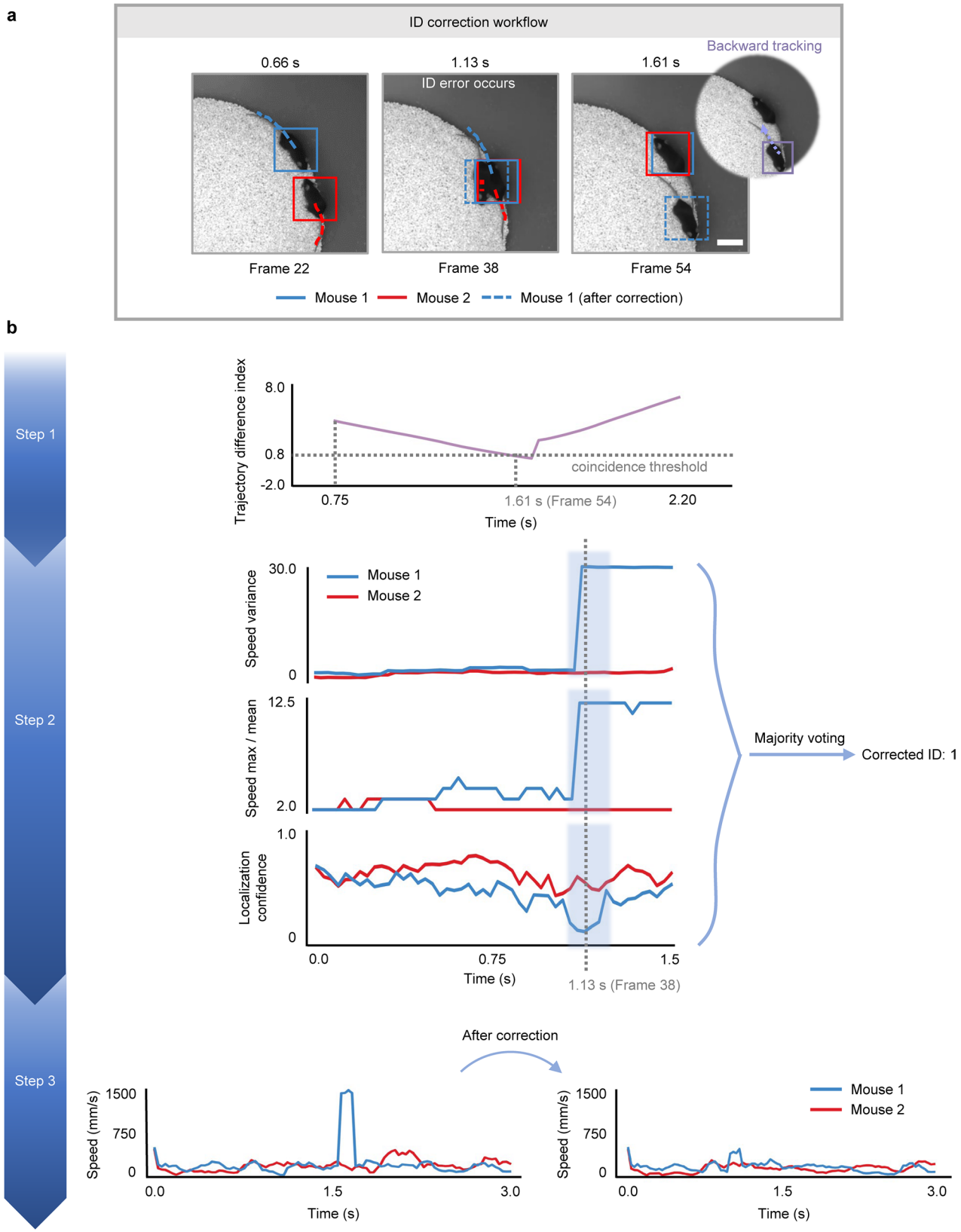
Extended Data Fig. 1 | Workflow of UDMT. a, Unsupervised tracking workflow. First, users create a project, and the first frame is segmented semi-automatically by a fine-tuned foundation model[28] to identify all animals to be tracked. Users can optionally refine the segmentation results with minimal manual input. Second, parameters will be optimized and training dataset will be constructed automatically. Third, a fine-tuned model will be trained in an unsupervised manner. Finally, the tracking results will be obtained after network inference, visualization and proofreading. **b**, Screenshot of the interface for initialization. **c**, Screenshot of the interface for tracking. **d**, Screenshot of the interface for proofreading.

https://doi.org/10.1038/s41592-026-03051-8



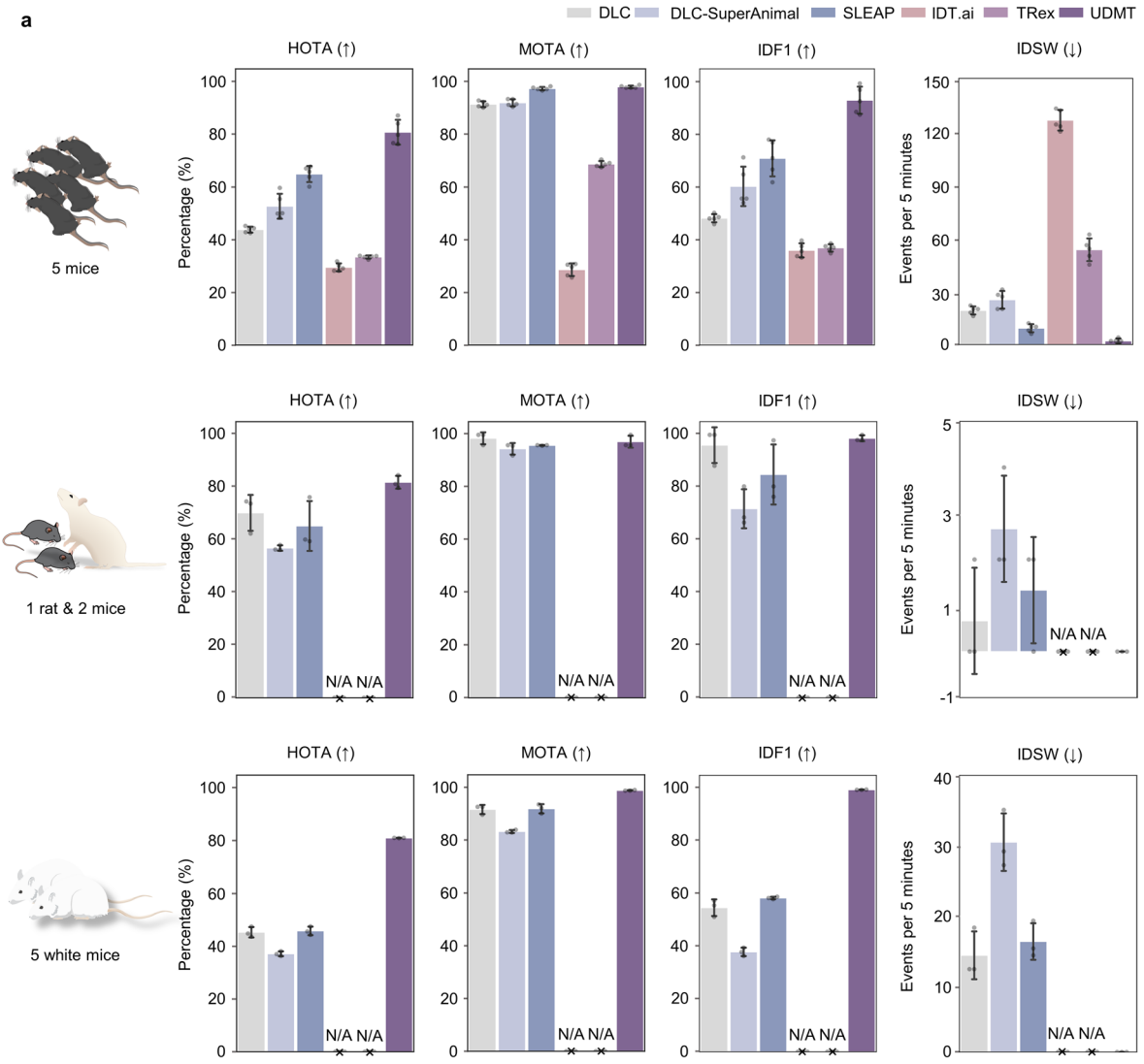
Extended Data Fig. 2 | Automatic parameter tuning module of UDMT. a, Flowchart of the automatic parameter tuning module. **b**, Tracking performance of UDMT with different search region sizes. The 5-mouse dataset (30 Hz frame rate, 18,000 frames) was used for quantitative evaluation. The search region size was calculated based on the search region scale and target size bias. Detailed information is provided in Supplementary Table 1. Lines indicate mean values and error bars indicate standard deviation. Gray dots represent individual samples (n = 5 video segments).

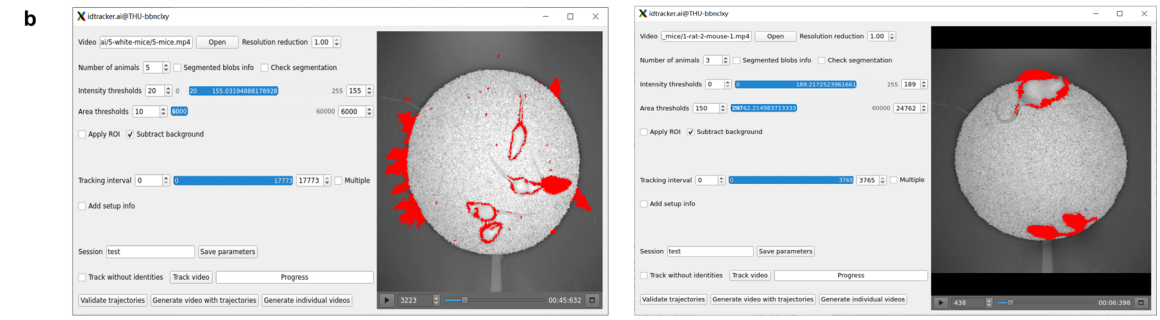
https://doi.org/10.1038/s41592-026-03051-8

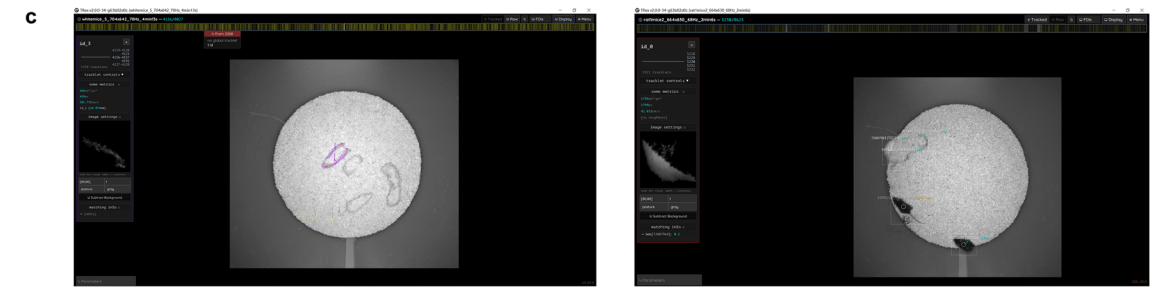


Extended Data Fig. 3 | ID correction module of UDMT. a, Three key frames illustrating the principle of ID correction. The lost target (mouse 1) in frame 54 is relocated and the correct ID is reassigned. Scale bar, 50 mm. **b**, Detailed ID correction workflow. First, the frame containing lost target is detected if the trajectory difference index of two mice falls below the threshold[56]. Second, UDMT detects an abrupt jump in movement speed (based on variance and max/mean) and a sudden drop in localization confidence to determine which individual (mouse 1 in this example) is assigned a wrong ID. Finally, the trajectory of the misidentified mouse was fixed using backward tracking. Line plots show the speeds of two mice as a function of time before and after ID correction.

https://doi.org/10.1038/s41592-026-03051-8





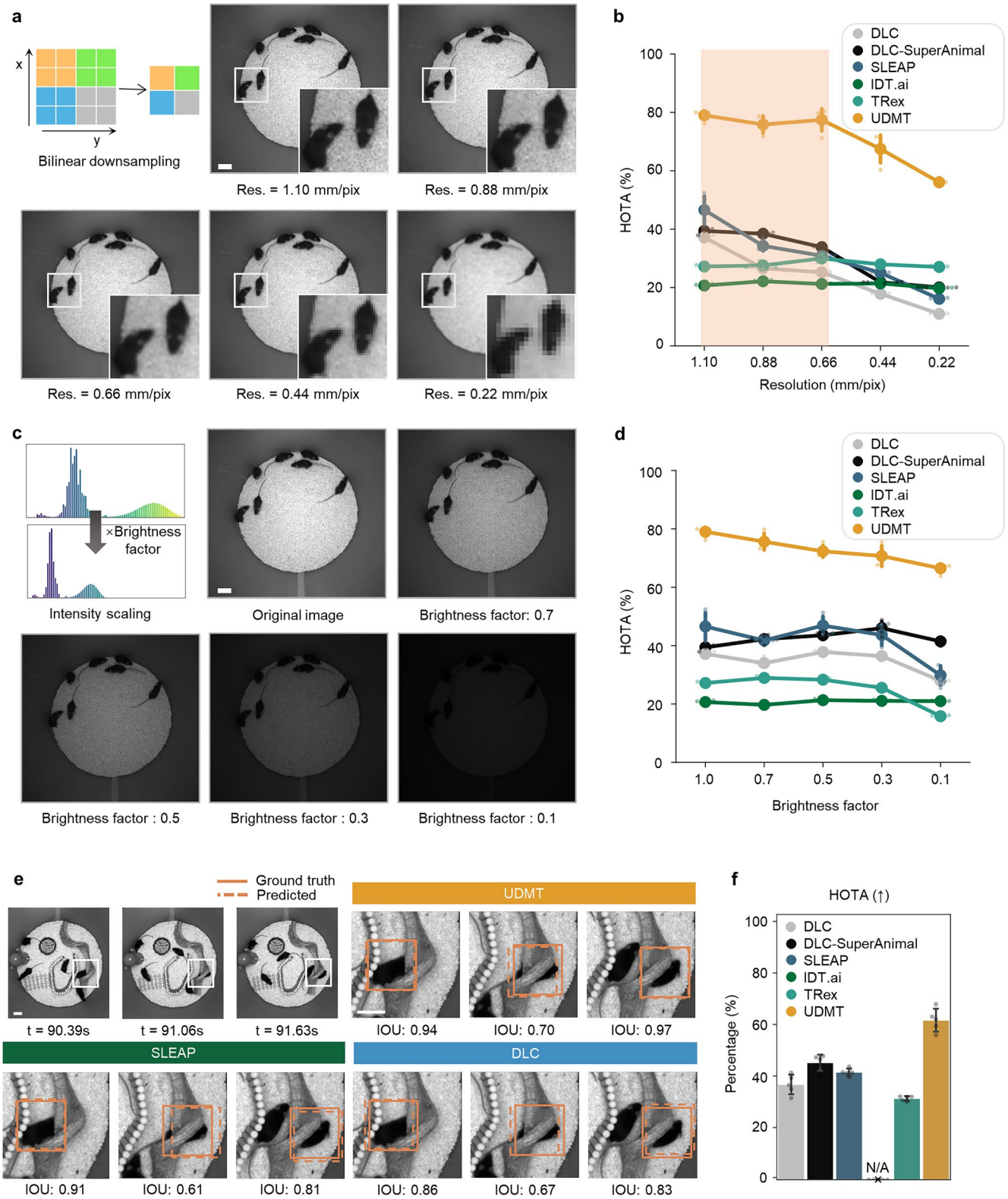


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41592-026-03051-8

Extended Data Fig. 4 | Comparing tracking performance of various methods on rodents. a, Tracking performance of UDMT, DLC, DLC-SuperAnimal, SLEAP, IDT.ai and TRex, quantified by HOTA, MOTA, IDF1, and IDSW. Bars indicate mean values and error bars indicate standard deviation. Gray dots represent individual samples. The 7-mouse dataset (top row, 67 Hz frame rate, 29,550 frames, n = 5 video segments), rat-and-mouse dataset (middle row, 68 Hz frame rate, 8630 frames, n = 3 video segments) and 5-white-mouse dataset (bottom row, 70 Hz frame rate, 17,760 frames, n = 3 video segments) were used for quantitative evaluation. N/A means that IDT.ai and TRex fail to track animals on the rat-and-mouse dataset and 5-white-mouse dataset. **b**, Examples showing how IDT.ai fails to segment low-contrast animals. **c**, Examples showing how TRex fails to segment low-contrast animals. Left, white mice. Right, white rat.

https://doi.org/10.1038/s41592-026-03051-8

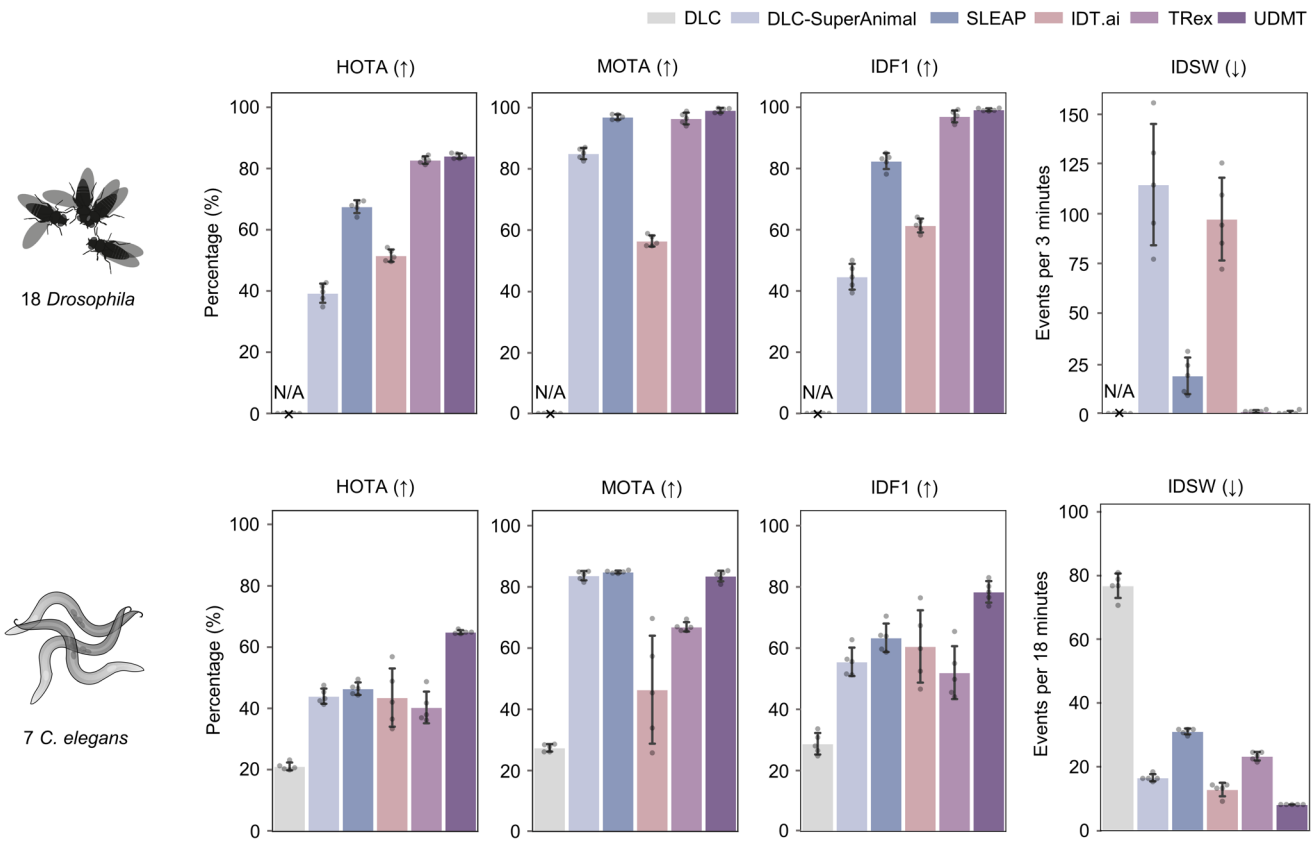


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41592-026-03051-8

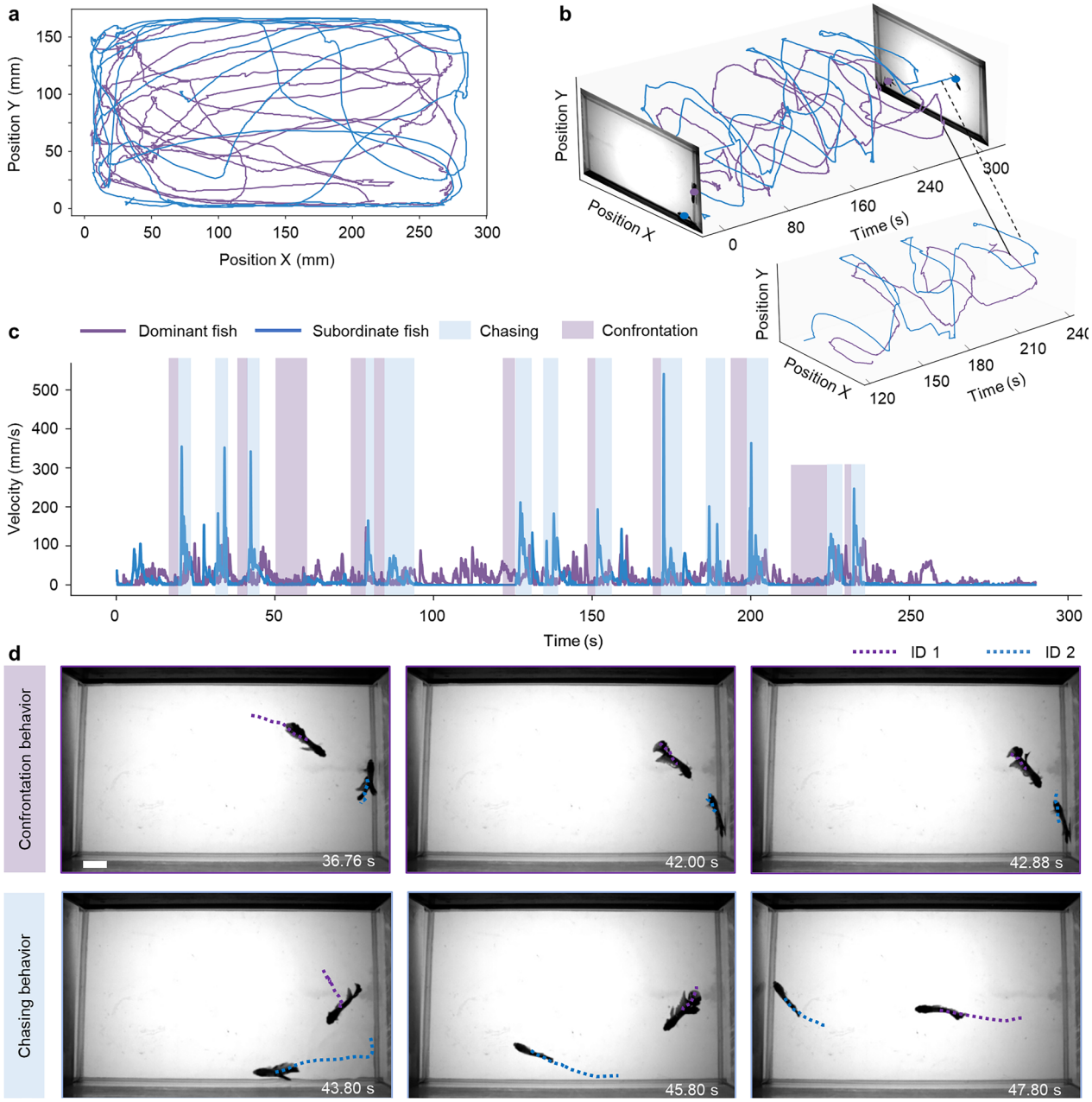
Extended Data Fig. 5 | Comparing UDMT with other methods on different recording resolution, brightness, and background complexity. The 7-mouse dataset (67 Hz frame rate, 29,550 frames) was used for quantitative evaluation. **a**, Representative video frames of different resolutions. Magnified views of the boxed regions are shown at the bottom of each image. Videos of different resolutions were obtained by downsampling the original high-resolution videos with bilinear interpolation. **b**, Quantitative relationship between image resolution and tracking accuracy of UDMT and baseline methods. Lines indicate mean values and error bars indicate standard deviation. Translucent dots represent individual samples (n = 5 video segments per dataset). The orange shaded area in the line plot indicates the range in which the performance of UDMT is robust to resolution degradation. **c**, Representative video frames under different brightness conditions. Videos of different brightness were obtained by multiplying the original bright video by different brightness factors (1.0, 0.7, 0.5, 0.3, 0.1) and rounding it into integers. **d**, Quantitative relationship between video brightness and tracking accuracy of UDMT and baseline methods. Lines indicate mean values and error bars indicate standard deviation. Translucent dots represent individual samples (n = 5 video segments per dataset). Baseline methods include DLC, DLC-SuperAnimal, SLEAP, IDT.ai and TRex. Scale bars, 50 mm. **e**, Tracking performance in complex environments. Representative frames and IOU metrics of UDMT, DLC and SLEAP are shown. Scale bars, 50 mm for all images. **f**, Quantifying the performance of UDMT, DLC, DLC-SuperAnimal, SLEAP, IDT.ai and TRex with HOTA. Bars indicate mean HOTA values across recordings, and error bars indicate standard deviation. Gray dots represent individual recordings. The 5-mouse dataset with complex environments (82 Hz frame rate, 35,728 frames, n = 5 video segments) was used for quantitative evaluation. N/A means that IDT.ai fails to track animals on this dataset.

https://doi.org/10.1038/s41592-026-03051-8



Extended Data Fig. 6 | Comparing tracking performance of various methods on *Drosophila* and *C. elegans* datasets. Tracking performance of UDMT, DLC, DLC-SuperAnimal, IDT.ai, SLEAP and TRex, quantified by HOTA, MOTA, IDF1, and IDSW. Bars indicate mean values and error bars indicate standard deviation. Gray dots represent individual samples (n = 5 video segments per dataset). The 18-*Drosophila*-2 dataset (54 Hz frame rate, 32,370 frames) and 7-*C. elegans* dataset (10 Hz frame rate, 8,630 frames) were used for quantitative evaluation. N/A means that DLC fails to track animals on the *Drosophila* dataset.

https://doi.org/10.1038/s41592-026-03051-8



Extended Data Fig. 7 | Behavioral analysis of two *Betta splendens* using UDMT. The interaction of two *Betta splendens* (betta fish) in the same arena is continuously recorded (51 Hz frame rate, 15,020 frames). Their swimming trajectories were extracted using UDMT. **a**, Projected trajectories of the two betta fish during the entire recording period. **b**, 3D (x-y-t) trajectories of the two betta fish in a 300-second time window. The trajectories during chasing are shown separately in the inset. **c**, Velocity of the two betta fish. Traces were averaged over 10 frames. The purple shaded area indicates the time of confrontation, during which the speed of both animals decreased to nearly zero. The blue shaded area indicates the time of chasing, when the speed of the escaping fish (ID 2) increased rapidly. **d**, Representative video frames showing the confrontation and chasing behavior of the two betta fish. The tracks show the movement of the two betta fish over 1.5 s. Scale bar, 20 mm.

nature portfolio

Corresponding author(s): Xinyang Li, Ziwei Li, Jiamin Wu, Qionghai Dai

Last updated by author(s): Jan 21, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The video acquisition was accomplished using MindVision V2.0.
Data analysis	Videos were tracked using UDMT V1.1.1. The complete code for UDMT is publicly available at https://github.com/cabooster/UDMT. All data processing and analysis were conducted using Python (3.8.0) scripts and our customized MATLAB (MATLAB 2021a, MathWorks) scripts. Deep learning models reported in this work were implemented with standard libraries in Python (3.8.0) and PyTorch (1.7.1, Facebook). Raw calcium imaging data from the head-mounted miniaturized microscope were motion-corrected using the open-source NormCorre algorithm in non-rigid mode (https://github.com/flatironinstitute/NoRMCorre). The data were then processed using DeepWonder (https://github.com/yuanlong-o/Deep_widefield_cal_inferece/tree/main/Alternative/DeepWonder) to extract neuronal masks and calcium traces. Spike inference was performed using the open-source MLspike (https://github.com/MLspike).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All behavioral recordings used in this study have been archived and made publicly available at https://zenodo.org/records/14580256. The source data of neuroethological research of freely behaving mice, including behavioral recordings and synchronized calcium imaging data, are publicly available at https://zenodo.org/records/14586426. In addition to this, all data collected in this study has been uploaded to https://zenodo.org/records/14580391 (part #1) and https://zenodo.org/records/14586218 (part #2).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. Sample sizes for the behavioural tracking and calcium-imaging experiments were determined by the availability of animals and by feasible, sufficiently long recording durations under ethical and practical constraints. These datasets were deemed adequate for the quantitative analyses of tracking performance and neural activity in this study.
Data exclusions	No data were excluded for the analysis.
Replication	All attempts at replication were successful.
Randomization	Randomization was not applicable, as this study evaluates a tracking method and animals/recordings were not assigned to intervention groups.

Blinding	Blinding was not applicable, as this study evaluates a tracking method and no intervention groups were formed. Data processing and quantitative analyses were performed using predefined automated pipelines.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Mice (C57BL6/J, ~2.5-4 months, male), Mice (C57BL6/J, Ai148D, ~2.5-4 months, male), Rat (wild-type SD, male), Drosophila (wild-type Canton S, sex not determined), C. elegans (wild-type N2, hermaphrodites), Betta splendens (male). Mice were housed in standard cages with a maximum of 5 mice per cage. Cages were housed in an environment with a 12/12h reverse dark/light cycle, and ambient temperature of 72°F and an ambient humidity of ~30%. Mice were provided food and water.
Wild animals	Not involved in this study.
Reporting on sex	The mice, rats, and Betta splendens used in this project are male. The biological sex of the Drosophila used in the study is unknown, but it is likely to be evenly distributed between females and males. The C. elegans are hermaphrodites.
Field-collected samples	Not involved in this study.
Ethics oversight	Animal protocol procedures were reviewed and approved by the Institutional Animal Care and Use Committee office of Tsinghua University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*