

Social Behavior Atlas: A few-shot learning framework for multi-animal 3D social pose estimation, identification, and behavior embedding

Pengfei Wei (✉ pf.wei@siat.ac.cn)

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences <https://orcid.org/0000-0003-1845-8856>

Yaning Han

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences <https://orcid.org/0000-0002-1650-2262>

Ke Chen

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Yunke Wang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Wenhao Liu

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Zhouwei Wang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Xiaoqing Wang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Chuanliang Han

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Jiahui Liao

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Kang Huang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Shengyuan Cai

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Yiting Huang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Nan Wang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Jinxu Li

Kunming Institute of Zoology, Chinese Academy of Sciences

Yangwangzi Song

Kunming Institute of Zoology, Chinese Academy of Sciences

Jing Li

Kunming Police Dog Base of the Chinese Ministry of Public Security

Guo-Dong Wang

Chinese Academy of Sciences <https://orcid.org/0000-0002-9407-4363>

Liping Wang

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Ya-Ping Zhang

Kunming Institute of Zoology <https://orcid.org/0000-0002-5401-1114>

Article

Keywords:

Posted Date: July 25th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3020951/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Machine Intelligence on January 8th, 2024. See the published version at <https://doi.org/10.1038/s42256-023-00776-5>.

**Social Behavior Atlas: A few-shot learning framework for multi-animal 3D social pose
estimation, identification, and behavior embedding**

Yaning Han^{1, 2, 3, 4, #}, Ke Chen^{1, 2, 3, 4, #}, Yunke Wang^{1, 3, 4, #}, Wenhao Liu^{1, 3, 4, 5}, Zhouwei Wang^{1, 2, 3, 4}, Xiaojing Wang^{1, 3, 4, 6}, Chuanliang Han^{1, 3, 4}, Jiahui Liao^{1, 3, 4, 7}, Kang Huang^{1, 2, 3, 4}, Shengyuan Cai^{1, 3, 4}, Yiting Huang^{1, 3, 4}, Nan Wang^{1, 2, 3, 4}, Jinxiu Li⁸, Yangwangzi Song⁸, Jing Li⁹, Guodong Wang⁸, Liping Wang^{1, 3, 4}, Yaping Zhang⁸, and Pengfei Wei^{1, 3, 4, *}

¹Shenzhen Key Laboratory of Neuropsychiatric Modulation and Collaborative Innovation Center for Brain Science, Shenzhen-Hong Kong Institute of Brain Science, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³CAS Key Laboratory of Brain Connectome and Manipulation, the Brain Cognition and Brain Disease Institute, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

⁴Guangdong Provincial Key Laboratory of Brain Connectome and Behavior, the Brain Cognition and Brain Disease Institute, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

⁵Department of Neuroscience, City University of Hong Kong, Kowloon Tong, Hong Kong, China

⁶Department of Physical Education, China University of Geosciences, Beijing 100083, China

⁷School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China

⁸State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

24 ⁹Kunming Police Dog Base of the Chinese Ministry of Public Security, Kunming 650204, China

25 [#]The authors contributed equally to this paper.

26

27 **Correspondence to:**

28 *Pengfei Wei, Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced

29 Technology, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen, 518055, P.R.China. E-mail:

30 pf.wei@siat.ac.cn

31

32 **Abstract**

33 The fact that multi-animal behavior quantification is still technically challenging nowadays
34 greatly limits the accuracy and granularity of social behavior analysis. Data labeling of deep-learning-
35 based approaches can be incredibly laborious, especially when multiple animals closely interact with
36 each other, under which circumstances animal identity switching and body occlusions are likely to
37 happen. To overcome the challenge, we designed a novel framework - Social Behavior Atlas (SBeA)
38 and it shows promising results. SBeA utilizes a much smaller number of labeled frames for multi-
39 animal 3D pose estimation, achieves label-free identification recognition, and successfully applies
40 unsupervised dynamic learning for social behavior classification. Our results also demonstrate that
41 SBeA is capable of achieving high performance across various species using existing customized
42 datasets. Therefore, we can build a multi-animal behavior atlas. To start with, we construct a social
43 behavior atlas for autism spectrum disorder (ASD) knockout mice, and SBeA reveals several
44 behavioral biomarkers which were neglected before. These findings highlight the immense potential
45 of SBeA for quantifying subtle social behaviors and provide a new paradigm in the fields of
46 neuroscience and ecology.

Introduction

Animal modeling is instrumental in the studies of human social disorders. But failures to capture their specific behavioral biomarkers impede our understanding¹. The biggest challenge to the quantitative analysis of animal social behavior is intraspecific physical resemblance². Although recent advances in deep learning-based multi-animal pose estimation^{3,4}, identity recognition³⁻⁶, and behavior classification⁷ have shown exhilarating performance in animal social behavior analysis, their application across various experimental scenarios is greatly compromised by the availability of high-quality benchmark datasets^{2-5,8}.

The model's performance of multi-animal pose estimation is decided by the number of labeled frames⁴. Although there are several well-annotated datasets for multi-animal pose estimation^{3,4}, they cannot cover the applications of diverse social behavior test paradigms. Additionally, the frequent occlusion of multiple animals is a challenge for manual data annotations. The performance of models would decrease because the manual labels of occluded frames are not precise. Combining multi-view camera array with 3D reconstruction technology can improve the pose estimation precision when facing frequent occlusions¹⁰, but these methods are designed for a single animal except for multiple animals^{9,10}.

Although image-based animal identification methods show advances in recognition performance, the precision of animal identification is restricted by data annotations^{5,6}. Animals have similar appearances, making it difficult to distinguish their identities when annotating identity datasets⁵, and adding new animals for identification requires additional data annotation^{5,11}. Unsupervised tracking-based methods are the alternative solutions to animal identification^{3,4}. They demonstrate high identification performance when the animals are a relatively long distance away from each other, but

the close interaction of animals can cause the identity swap problem². The frequent close interaction results in these methods cannot maintain identities for a long period of time².

New abnormal social behavior patterns from animal disease models cannot be covered by existing behavioral classification datasets. Some sub-second behavioral modules are casually ignored in behavior labeling because they are too short¹⁰. The two reasons cause supervised behavior classification methods are not suitable for detecting unusual and short behaviors⁵. Recent advances in unsupervised behavior decomposition and embedding methods are appropriate to reveal subtle behavioral differences^{10,12}, but they are only designed for a single animal. AlphaTracker is designed for the unsupervised clustering of social behavior using human-defined features⁸. These features cannot distinguish the subtle behavior interactions constructed by limbs and paws. The decomposition and embedding of multi-animal social interactions should follow the natural structure of their behaviors¹⁰.

To address these challenges, we propose SBeA, a few-shot learning framework for multi-animal 3D pose estimation, identity recognition, and social behavior classification. We propose a continuous occlusion copy-and-paste algorithm (COCA) for data augmentation in SBeA, combined with a multi-view camera, to achieve multi-animal 3D social pose estimation with a few data annotations (~400 frames)^{13,14}. We propose a bidirectional transfer learning identity recognition strategy, achieving zero-shot annotation of multi-animal identity recognition with an accuracy rate exceeding 90%¹⁵⁻¹⁷. We extend Behavior Atlas, an unsupervised dynamic behavior decomposition framework, from a single animal to multiple animals, which achieves unsupervised fine-grained social behavior module clustering with a purity exceeding 80%^{10,18,19}. In the study of free social behavior between autism model animals and normal animals, the application of SBeA enables automatically identifying of

91 animals with social abnormalities and explores the precise characteristics of these abnormal social
92 behaviors. This demonstrates the excellent performance of SBeA as a quantitative tool for studying
93 animal social behavior. The models and frameworks developed for SBeA can be also applied to mice,
94 parrots, and Belgian *Malinois*, showcasing its strong generalization abilities suitable for various
95 application scenarios. We also provide a user interface to facilitate users without coding backgrounds
96 to apply SBeA in their own social behavior experiments.

97

98

Results

SBeA: from multi-animal markerless 3D pose tracking to unsupervised social behavior mapping.

The focus of SBeA is on quantifying the behavior of freely social animals comprehensively. It presents two significant challenges: pose tracking and behavior mapping. Pose tracking involves identifying the key body parts of each animal as well as their identities, which is particularly challenging when dealing with animals that look similar². To address this issue, a novel free social behavior test paradigm has been developed that involves a multi-view camera array (Fig. 1a). This approach captures the animals covering more view angles and helps to overcome the challenge of frequent occlusion^{9,10,18,19}. The camera array is used to capture images of a checkerboard for camera calibration, followed by videos of two free-moving animals for the social behavior test (Video capture phase 1, Fig. 1a). Finally, the array captures videos of single free-moving animals to facilitate animal identification without the need for human intervention (Video capture phase 2, Fig. 1a).

After the video acquisition, the multi-animal contour of video capture phase 1 and the single-animal pose of video capture phase 2 are manually annotated for the training of AI to output the 3D poses with identities of animals (Fig. 1b and c). The design of this AI model was separated into four stages for function integration. Through these multistage networks, the task of multi-animal video instance segmentation, pose estimation, and identity recognition was achieved with a relatively small number of manual annotations (~400 frames), as shown in Fig. 1d (left). By incorporating camera parameters, the above results from various camera angles were combined and matched based on geometric constraints to reconstruct 3D pose trajectories with identities for each animal (Fig. 1d, center and right).

After conducting pose tracking, the process of behavior mapping involves breaking down the

trajectories of animals into distinct behavior modules, and then, using appropriate metrics to obtain a low-dimensional representation of these modules¹⁰. In the context of SBeA, the framework for decomposing behavior is extended from a single animal to multiple animals¹⁰, with their 3D trajectories being separately decomposed into locomotion, non-locomotor movement, and body distance components (Fig. 1e top and middle). These parallel components are then divided into segments and subsequently merged into social behavioral modules using the dynamic behavior metric (Fig. 1e bottom). Overall, this process utilizes a nature-inspired structure for behavior decomposition and provides a dynamic approach to understanding social behaviors in groups of animals.

To gain insight into the distribution of features within social behavioral modules, it is necessary to convert them into low-dimensional representations (Fig. 1f). These representations incorporate both spatial and temporal aspects, with the spatial aspect being captured by low-dimensional embeddings of distance features in the SBeA framework (Fig. 1f left). The temporal aspect is represented by the social ethogram (Fig. 1f right). In SBeA, social behavioral modules are first clustered based on their spatial characteristics and then expanded into the temporal dimension to construct the social ethogram. This approach allows for a more comprehensive understanding of the distribution of features within social behavioral modules.

137

A general augmentor to extend single animal pose estimation to multiple animals of SBeA.

The use of deep learning for social pose estimation has been beneficial in enhancing the acquisition of data for multiple body parts in animals, as previously demonstrated in literature^{3,4}. However, due to the flexible social interactions among animals, creating a comprehensive training dataset for deep learning-based social pose estimation is a challenge. Inadequately trained deep neural

143 networks tend to produce higher tracking errors, particularly in frames with close animal interactions
144 ². To address this issue, we introduce a novel animal tracking method, COCA, (Fig. 2a) in our SBeA
145 framework. Previous studies show that image copy-paste can increase the precision of instance
146 segmentation and multi-object tracking^{13,14}. Here for multi-animal tracking, we extend the above
147 methods to COCA, which generates the virtual scenario with instance occlusion. By capturing a short
148 video of multiple animals behaving freely in an open field, SBeA obtains sufficient elements
149 (background and animal instances) to generate the virtual dataset. These elements synthesize the
150 complex interactive relationships between animals without the need for manual annotations, resulting
151 in a sufficiently large virtual dataset to train deep neural networks.

152 During free behavior, it is common for animals to overlap, leading to loss of tracking in single-
153 view cameras. To address this, SBeA employs a multi-view camera array to capture video streams,
154 enabling compensation for the visual field of the cameras and facilitating continuous tracking (Fig. 2b)
155 ^{9,10,19}. Background and trajectories can be extracted through background subtraction algorithms applied
156 to the raw video streams (Fig. 2c left top and left middle). In addition, frames with close social
157 interactions can be extracted for manual contour annotations (Fig. 2c left bottom). A lightweight
158 instance segmentation deep neural network, YOLACT++, is trained with self-training using
159 approximately 400-800 annotated contour frames (Fig. 2c center bottom), which enhances its
160 performance while ensuring time-efficiency^{20,21}. The well-trained YOLACT++ predicts animal masks
161 of video streams, and animal instances can be cropped based on these masks. As some trajectories of
162 multiple animals can overlap in the same spatial position across different periods, merging animal
163 instances, backgrounds, trajectories, and masks can generate virtual scenarios with various occlusion
164 relationships and mask labels (Fig. 2c center top and center middle). The COCA increases the scale of

165 the training dataset without additional manual annotations, producing a VIS dataset with successive
166 frames of behaving animals and annotations. To capture the spatial-temporal patterns of occluded
167 animals, the video instance segmentation with transformers (VisTR) method is modified and applied
168 to the virtual VIS dataset as it can segment instances at the sequence level (Fig. 2c right top)²². Well-
169 trained VisTR can patch the raw video streams to display only one animal in each video (Fig. 2d left
170 top and left middle). Thus, pose estimation models trained for single animals can be used to predict
171 single animal poses on these patched videos using COCA as a general augmentor (Fig. 2c right bottom,
172 and 2d left bottom). Finally, the single-animal poses of each patched frame are merged into multi-
173 animal poses (Fig. 2d left top, left middle, and left bottom).

174 The subsequent step is the 3D reconstruction (Fig. 2e). Firstly, the MouseVenue3D automatic
175 camera calibration system is employed to acquire the camera parameters of the camera array (Fig. 2e
176 left top)^{9,19}. Then, based on the epipolar constraint of camera parameters, the combination of each
177 animal instance in each camera view is optimized to achieve minimum reprojection error (Fig. 2e left
178 bottom). The optimized 3D skeletons of the single frame in Fig. 2d right bottom are presented in Fig.
179 2e right top and bottom. In the 3D skeleton, the close contact between two animals can be quantified
180 (Fig. 2e right top and bottom).

181 Compared with the square increasing of multi-animal pose estimation methods such as maDLC,
182 the pose annotation strategy in SBeA is linearly increasing with body points and the number of animals
183 (Fig. 2f)³. Considering the diversiform social interaction of animals, routine multi-animal pose
184 estimation methods need to annotate more data on frames with various social interactions to get higher
185 precision.

186 In order to compare the tracking performance of SBeA with maDLC, we create a well-annotated

dataset Social Black Mice for Video Instance Segmentation (SBM-VIS). According to the distance distribution of the test dataset, the clustering algorithm is used to separate close interaction (Fig. 2g, the left orange stem) and other conditions. The pixel root-mean-square error (RMSE) of all data is significantly lower than the close interaction of about 2 pixels of different body parts (Fig. 2h). But compared with multi-animal DeepLabCut (maDLC) and SLEAP, SBeA still has significantly lower RMSE of animal close interaction, with 800 pose-labeled frames are used to train maDLC/SLEAP and 400 pose- and 400 mask-labeled frames are used to train SBeA (Extended Data Fig. 1, and Extended Data Fig. 2). For all of the test data, SBeA has significantly lower RMSE than maDLC in the Nose, Left ear, Right ear, Root tail, Mid tail, and Tip tail while maDLC has significantly lower pixel RMSE in the Back, Right front paw and Left front paw (Extended Data Fig. 1a). SBeA has significantly lower RMSE than SLEAP except for the Back and Right hind limb (Extended Data Fig. 2a). For the close contact part, most of the RMSE of SBeA are significantly lower than maDLC (Extended Data Fig. 1b), and SBeA has significantly lower RMSE than SLEAP except for the Neck (Extended Data Fig. 2b). These results show that SBeA can get higher precision with fewer manual annotations than routine multi-animal pose estimation methods such as maDLC and SLEAP.

SBeA needs no data annotations for multi-animal identification.

Accurately distinguishing the identities of free-moving animals is crucial for social behavior tests, particularly in studying treatment-induced behaviors in transgenic animal models^{10,23,24}. However, frequent occlusion of these animals can lead to imprecise identification even with physical markers. Moreover, the animals are the same breed to reduce the influence of irrelevant experiment variables with indistinguishable appearances for human annotators. It causes difficulties in creating the animal

209 identification dataset to train deep neural networks such as SIPEC⁵. To address these challenges, we
210 propose bidirectional transfer learning in animal identification as a solution in SBeA (Fig. 3a). Transfer
211 learning allows artificial neural networks to use previous knowledge in new tasks¹⁶. For animal
212 segmentation and identification tasks, the knowledge between them can be shared bidirectionally with
213 each other. So, the segmentation model trained for multi-animals can be transferred to single-animal
214 segmentation, and the identification model trained for single-animals can be transferred to multi-
215 animal identification. The bidirectional transfer learning of them avoids unnecessary manual data
216 annotations.

217 Well-trained VisTR in Fig. 2 can be used to segment single-animal instances from multiple view
218 angles (Fig. 3b). These instances are then cropped, cascaded, and resized to generate training data for
219 an identification model based on EfficientNet (Fig. 3c, left and center)²⁵. After that, LayerCAM is used
220 to evaluate the patterns for identification recognition (Fig. 3c right)²⁶. Before using the identification
221 model in multi-animal instances, the cascaded and resized image frames were prepared (Fig. 3d, right).
222 By using the best geometric constraint of 3D poses, instances from each frame view angle of each
223 animal were matched to construct input frames of the identification model (Fig. 3d, left). Finally, the
224 well-trained model outputted the top prediction probabilities to append the identities of instances and
225 3D poses. LayerCAM was also employed to verify the recognition patterns for identification (Fig. 3e).

226 To evaluate the identification performance of SBeA, we conducted experiments with ten
227 C57BL/6J mice having tail markers, where we recorded their free behaviors for 5 minutes. The tail
228 markers are convenient for experimenters to distinguish the identities of each mouse in videos. The
229 first 4 minutes of data were used for training the identification model, and the last 1 minute was used
230 for validation. The confusion matrix of the validation data demonstrated that the model can identify

most of the mice (Fig. 3f). The t-SNE algorithm was used to create a 2D feature representation of the identified mice (Fig. 3j). The features of mice with ID M4 and M5 were found to be mixed with other classes, as quantified by the silhouette coefficient (Fig. 3h). The sorted validation precision of the identified mice showed that the mouse with ID M4 had the lowest precision of approximately 0.4 (Fig. 3i).

To assess the identification model's performance in multi-animal data, we recorded the free social behaviors of 5 paired C57BL/6J mice identified by SBeA for 15 minutes. We manually verified the identities of mask reprojection images and 3D poses frame by frame (Fig. 3j). The results indicated that although some of the single mouse identity precisions were lower (Fig. 3i), the overall precision in identifying pairs of mice could be higher than 0.85, as seen in the case of the pairs of M3&M4 and M5&M6.

We also investigate if the number of animals would influence the identification recognition precision. Previous research suggests that the identification precision may decrease with an increasing number of animals^{6,27}. We increased the amount of training data to counteract the precision decrease. Our results indicate that for a group of 22 mice, a 15-minute video recording can achieve similar precision to that of 10 mice with a 5-minute recording (Fig. 3k). These findings have important implications for optimizing study design and ensuring accurate identification of individual animals in social behavior experiments.

Our research has revealed that the precision of animal identification can be influenced by the experiment apparatus (Fig. 3l). Specifically, we found that open fields with different diameters (50 cm and 20 cm) can impact the precision of animal identification conducted on the same ten mice. The results indicate that the precision of identification in the 20 cm open field is significantly higher than

that in the 50 cm field (Fig. 3l). This difference may be due to the higher pixels per inch (PPI) of mice.

Further, we tested the stability of identification patterns. Animals would groom themselves, which could change the patterns of identities¹¹. We compared the identification precision of two separate groups of mice. One group underwent both identity video recording and social behavior tests on the same day, while the other group underwent social behavior tests one week after their identity videos were recorded (Fig. 3m). We manually verified the identities of mask reprojection images and 3D poses frame-by-frame. Our analysis revealed that while there was no significant difference in the precision of mask reprojection images between the two groups, the precision of 3D poses in the group that underwent social behavior tests one week after the recording of their identity videos was significantly lower than that of the group that underwent both on the same day (Fig. 3m). The observed decrease in precision of 3D poses indicates a decline in identification precision. Shorter intervals between the recording of identity videos and social behavior tests could enhance the accuracy of identification recognition.

We evaluated the feature intensity of the identification model used to distinguish different animals at last in this chapter. We used open fields with diameters of 50 and 20 cm, respectively, in which the same 10 mice were allowed to freely engage in social behavior with 2 mice per trial. The pose tracking point Root tail with tail markers was used as a control against other body parts (Fig. 3n). We calculated LayerCAM values to quantify the feature intensity of each body point. The results showed that the Root tail in the 50 cm group had more significant feature intensities than in the 20 cm group. This finding suggests that a higher PPI can enable the identification model to capture more available fur pattern features and thereby overcome errors resulting from marker occlusion. In contrast, the identification using low animal PPI requires the use of stronger markers to maintain sufficient

recognition precision.

SBeA reveals the social behavioral structure in the atlas by unsupervised machine learning.

Following pose tracking, it is necessary to map the trajectories with animal identities to a low-dimensional space to gain insights into behavior (Fig. 4a). Recent research has indicated that the body language of social animals can be represented through sequential behavioral motifs or modules²⁸. Thus, we expand our prior work on the animal behavior mapping framework to encompass multiple animals, Behavior Atlas (BeA), which was initially developed for a single animal. The concepts of parallel and dynamic behavior decomposition from BeA have been adopted in our new framework SBeA (Fig. 4b and c). In the social process, the distance between animals is an essential component²⁹. Hence, in addition to using non-locomotor movement to assess body movement and locomotion to evaluate body displacement, body distance is also utilized to evaluate the relationships of body position (Fig. 4b). After parallel decomposition, each component is decomposed further using dynamic time alignment kernel (DTAK) to retain the natural dynamic structures of behavior (Fig. 4c). To distinguish subtle structures of social behavior, the temporal points of decomposition for each component are merged through logical addition (Fig. 4d). The aforementioned steps enable the metric of social behavior, resulting in the transformation of continuous pose trajectories into discrete social behavior modules.

Then, the social behavior modules are embedded in a low-dimensional space for behavior representation (Fig. 4e and f). All of the social behavior modules from different experimental trials need to be represented in a common feature space. That induces two questions: the first is what feature is reasonable to represent social behavior in a low-dimensional space, and the second is how to create a common feature space under the big behavioral data^{30,31}. For the first question, the distance

component is chosen for the feature representation of social behavior modules (Fig. 4e left). The dimensionally reduced distance component by uniform manifold approximation and projection (UMAP) is beneficial to improve the separation of behavior atlas verified by previous studies^{9,10,18,19,32}. But with the increase of data scale, the computational consumption of UMAP would be unacceptable because of limited memory space, which is the second question. To solve the second question, the residual multilayer perceptron (ResMLP) is combined with UMAP for feature representation (Fig. 4e right)³³. A part of the social behavior feature frames is extracted randomly to build up the feature representation of distance dynamics by the UMAP. Then, the mapping from extracted social behavior feature frames to distance dynamics is trained by ResMLP for the feature encoding. Further, the rest of the social behavior feature frames are decoded by ResMLP to distance dynamics. The distance dynamics are embedded by DTAK and UMAP to construct the social behavior atlas (Fig. 4f). To reveal the distributions of different social behavior modules, based on density clustering, we modified the watershed algorithm to automatically determine the best cluster density with upper and lower boundaries. At last, the social behavior modules of the same clusters are manually identified and defined (Fig. 4g).

In constructing the social behavior atlas, the inclusion of the distance component is crucial. By using the distance component, the social behavior atlas can maintain the overall body distance structures of social behavior modules (Fig. 4h left), while the absence of the distance component leads to a lack of observable patterns in the distribution of distance (Fig. 4h right). To compare the effectiveness of the distance representations in the atlases, the map/body distance metric is utilized, with higher values indicating better performance in distance representation (Fig. 4i). Results show that the distance component is essential in achieving a high map/body distance, indicating the importance

of including this component in constructing the social behavior atlas. Additionally, the $0.45 \pm 0.32s$ temporal duration of merged behavioral modules reveals that the SBeA framework can effectively decompose social behavior into dynamic sub-second motifs (Fig. 4j)^{10,34}. The ResMLP can address issues related to the memory cost of large behavioral data while also reducing computational time consumption compared to using UMAP alone. More than 5000 frames can get time benefits from ResMLP, and the time benefits will increase with the number of frames (Fig. 4k). Then, the time proportion of identified behavioral modules is quantified to evaluate their temporal precision (Fig. 4l). The time proportion of the typical social behavior such as allogrooming conforms to previous studies on social behavior³⁵. Further, the feature correlations between the intra- and inter-clusters of each social behavior class are compared for the evaluation of clustering consistency (Fig. 4m). The intra-feature correlations of each social behavior class are significantly higher than inter-feature correlations, and the intra-feature correlations distribute consistently near to 1, in turn, the inter-feature correlations distribute in the weak negative correlation. These unsupervised validation measures demonstrate the effectiveness of the SBeA framework in accurately mapping social behavior.

In addition to unsupervised validation, we conducted supervised validation of SBeA using the PAIR-R24M dataset (Fig. 4n)³⁶. The dataset provides 3D poses, social behavior labels, and subject behavior labels of rats in free social behavior. We used SBeA to construct the social behavior atlas for the dataset, and appended the three social labels (close, chase, and explore) to each behavior module (Fig. 4o). The distributions of the three social labels were separated and matched their similarity relationship. The 121 combinations of subject behavior labels also showed distribution patterns in the social behavior atlas (Fig. 4p). The social labels such as close and explore were consistent with the close distance distribution in the distance map, and the chase label was consistent with the distance

transition zone of the distance map (Fig. 4q). To quantify the clustering performance, we used the cluster purity of social and subject behavior labels (Fig. 4r and s). For the upper boundary of clustering, 14 classes were clustered with a mean cluster purity of 0.77 ± 0.16 (Fig. 4r). For the lower boundary of clustering, 405 classes were clustered, and the probability of cluster purities greater than 0.95 was significantly higher than for other purities (Fig. 4s). These results suggest that SBeA can classify the behavior clusters with high cluster purity.

SBeA identifies *Shank3B* knockout mice in free-social interactions by subtle behavior modules.

Social behavior can serve as an indicator of the genetic variations that underlie neuropsychiatric disorders³⁷. SBeA is well-suited for this purpose, as it allows for a detailed characterization of social behavior at an atlas-level. To test the ability of SBeA to detect genetic differences from social behavior, we utilized an animal model of autism spectrum disorder (ASD): *Shank3B* knockout mice^{10,23}. While abnormal individual behaviors of these mice have been previously identified, the limitations of existing techniques have made it difficult to fully understand their abnormal free social behaviors^{10,23}.

To distinguish between *Shank3B* knockout (KO) mice and wild-type (WT) mice, a free-social behavioral paradigm was designed based on the framework of SBeA, which consists of three steps: identity recording, social behavior recording, and SBeA processing (Fig. 5a). First, the home-caged WT and KO mice were randomly shuffled and recorded for 5 minutes each using the MouseVenue3D system to obtain identity information. After identity recording, the mice were randomly grouped into three pairs (WT-WT, WT-KO, and KO-KO) for social behavior recording, with each pair of mice recorded for 15 minutes. The identity and social behavior data were then processed using SBeA for 3D pose tracking and behavior mapping. The experiment used a total of 5 WT and 5 KO mice, resulting

363 in 45 unique pairs of mice, including 10 WT-WT, 10 KO-KO, and 25 WT-KO pairs. To ensure equal
364 representation of each group, the number of WT-KO pairs was reduced from 25 to 10 through random
365 sampling. Before behavior mapping, the raw trajectories were copied and switched to capture the
366 direction of social behavior between WT and KO mice. This resulted in a total of 60 pairs of trajectories
367 for behavior mapping using SBeA.

368 The social behavior atlas with distance map is shown in Fig. 5b. After the construction of the
369 social behavior atlas, the density map is calculated to compare the social behavior distribution of each
370 group by kernel density estimation (Fig. 5c). Density map shows obvious differences across the three
371 groups. Combining with the distance map, the WT-WT group shows social behavior phenotypes with
372 flexible distances from close to far, the KO-KO group shows more abnormal social behaviors than the
373 WT-WT group, and WT-KO shows more close social interaction than the WT-WT group.

374 The 260 social behavior modules identified in the social behavior atlas were clustered to reveal
375 their coincident patterns (Fig. 5d). Then the principal component analysis (PCA) was employed to
376 determine the percent variability explained by each principal component to compare the three groups
377 (Fig. 5e). The results indicated that three components could account for 90% of the variance, while 11
378 components could account for 99% of the variance. To construct the phenotype space of the three
379 groups, UMAP was used for dimensional reduction of the social behavior modules, with the dimension
380 number set to 3 based on the 90% variance explanation, owing to the more robust feature representation
381 of non-linear dimensional reduction (Fig. 5f). The distributions of the three groups in the phenotype
382 space were found to be segregated, matching the distribution of the density map (Fig. 5c).

383 Further, SBeA was utilized to identify subtle social behavior modules that distinguish KO and
384 WT mice. The two-way ANOVA was used to compare the behavior fractions between the three groups,

385 and 24 social behavior modules were found to have significant differences (Fig. 5h). To reduce the
386 redundancy of these results, angle spectrum clustering, which combines PCA and hierarchical
387 clustering, was proposed (Fig. 5g). The social behavior modules were merged based on their angular
388 separation of features, resulting in the identification of 9 social behaviors, as determined by human
389 analysis (Fig. 5i).

390 The 9 social behavior modules identified through SBeA highlighted significant differences among
391 the three groups. The WT-WT group exhibited more allogrooming, a prosocial behavior, than the WT-
392 KO and KO-KO groups³⁸. Conversely, allogrooming was rare in unstressed partners and even rarer in
393 Shank3B KO mice, suggesting an antisocial behavioral phenotype³⁵. The exploring behavior of the
394 WT-WT group was significantly higher than that of the KO-KO group, which displayed reduced motor
395 ability or social novelty^{10,23}. In the WT-KO group, social behavior with significant differences was
396 divided into two parts, namely, peer sniffing and independent grooming. Peer sniffing was observed
397 more frequently in the WT mouse, particularly when the KO mouse was grooming or in locomotion,
398 indicating a behavioral phenotype of curiosity. Furthermore, the KO mouse could induce higher
399 interest in the WT mouse than vice versa. Independent grooming could be an imitation of the WT
400 mouse by the KO mouse, and in the KO-KO groups, the higher incidence of independent grooming
401 could be attributed to the increased individual grooming of each mouse. In addition to increased
402 independent grooming, two abnormal behavior phenotypes, namely, synchronous behaviors and
403 immobility, were observed. The synchronous behaviors displayed 5 subtypes, including grooming,
404 hunching, rearing, sniffing, and micromovement, indicating greater behavior variability in free-social
405 conditions compared to individual spontaneous behavior of KO mice¹⁰. These findings demonstrate
406 that SBeA can differentiate genetic mutant animals based on social behavior and identify genetic

mutant-related subtle social behavior modules.

SBeA is robust to be used in different environments across species.

To assess the generalizability of SBeA to different animal species and experimental settings, the behaviors of birds and dogs were captured using the MouseVenue3D system with varying device configurations¹⁹. The animals were prepared to have as similar appearances as possible (Fig. 5a top and 5e top), and it was difficult for human experimenters to separate two animals from the randomly selected frames. The resulting videos were manually annotated to train the AI of the pose tracking component of SBeA (Fig. 6a bottom and 6e bottom), using 19 body parts for birds and 17 body parts for dogs, based on previous studies^{39,40} (Fig. 6b and f). Then, they were mapped to the social ethogram and behavior atlas (Fig. 6c and g). In total, 34 and 15 social behavior classes were identified for birds and dogs, respectively, and their typical cases were visualized in 3D (Fig. 6d and h). The 3D pose tracking of birds showed clear identification of their claw touching their rectrix, while the 3D pose tracking of dogs was robust to occlusion even in the lying posture.

To evaluate the performance of the SBeA algorithm in tracking birds and dogs, various metrics were employed (Fig. 6i-m). While dogs have a higher tracking likelihood than birds, both achieve a satisfactory level of tracking precision (Fig. 6i)¹⁰. The 3D reprojection error is significantly higher for dogs due to incomplete camera coverage (Fig. 6j). In terms of identity recognition, both birds and dogs have higher precision than mice (Fig. 6k). The results of feature correlation (FC) show that all of the intra-FC of clusters are significantly higher than inter-FC (Fig. 6l and m). But from the distribution of FCs, the clustering performance of birds is better than dogs. The feature mix-up of intra- and inter-clusters is influenced by the 3D pose tracking precision. The error of 3D pose tracking such as target

429 loss in dogs would degrade the performance of SBeA clustering. The LayerCAM analysis reveals no
430 significant differences in feature values between birds and dogs, except for the Mid back and Nose of
431 dogs, which may be attributed to the loss of nose detection in video captures (Fig. 6n and o). The
432 identification recognition network automatically reduces the feature weights on the body part with
433 target loss or occlusion to keep the higher recognition precision of identities. These results demonstrate
434 that SBeA is robust enough to be applied to different animal species in various experimental settings,
435 making it a versatile tool for the study of social behavior in animals.

Discussion

Here we have presented SBeA, a few-shot learning framework for 3D pose estimation, identification, and behavior embedding of multiple free-social animals. SBeA builds upon the BeA framework, extending it to enable multi-animal pose estimation and social behavior clustering^{9,10,18,19}. The method reduces the labor required for annotation compared to traditional approaches for pose estimation and identification³⁻⁵. By utilizing four cameras, SBeA overcomes the issue of occlusion and reconstructs 3D behaviors accurately. Notably, SBeA resolves the challenge of animal identification over extended time frames, facilitating the study of close social interactions². The technique is highly versatile and has been successfully applied to various animal species, including *Shank3B* knockout mice, where it revealed synchronous behaviors and reduced social interest. SBeA's cross-species application has been verified in birds and dogs. In summary, SBeA represents a breakthrough in deep learning-based pose estimation and identification, offering numerous potential applications in animal behavior research.

Although the benchmark datasets are critical to the advances in deep learning tools³, the large labeled data number could render them infeasible¹⁷. SBeA get rid of the dependency of large datasets and achieves only using hundreds of labeled frames to track the 3D poses and identities of multiple animals in millions of new frames. Recent studies have shown that the precision increasing of large Transformer models in human pose estimation^{41,42}, but the benchmark datasets of animals are still too small to apply them³. The data generation strategy in SBeA can be a bridge between small animal datasets and large models in current stage. The phenotypes of social behavior are diverse, which are difficult to be comprehensively predefined in a dataset^{10,36,43}. The unsupervised clustering in SBeA provides an unbiased way to classify undefined social behavior modules. It also facilitates the establish of a comprehensive social behavior dataset.

458 The maDLC and SLEAP are two excellent tools that can be applied to a variety of animal models^{3,4},
459 but they do not include the mechanism for maintaining animal identities during long-term experiments,
460 which influences the accuracy of building a behavioral representation relying on the identities of
461 animals². SBeA incorporates the identity recognition approach of idTracker.ai and TRex, utilizing deep
462 neural networks to directly learn the appearance features of animals^{6,44}. This results in the alleviating
463 of identity swap problem, which allows to filter the frames with higher error rates. Additionally, SBeA
464 provides an extension of 2D tracking tools to 3D movement tracking, which is critical for making
465 accurate inferences about animal behavior^{2,9,19}.

466 One potential area for future research to improve SBeA is the development of an end-to-end model
467 that can reduce storage consumption. To accomplish this, the process of data generation could be
468 incorporated into a video instance segmentation model. Additionally, the identity videos available in
469 this context may contain sufficient information to train a deep learning model for tasks such as multi-
470 animal segmentation, identification, and pose estimation. Furthermore, the behavior atlas of a single
471 animal could be combined with a social behavior atlas of multiple animals through an algorithmic
472 bridge from BeA to SBeA that facilitates not only social behavior analysis but also other forms of
473 analysis within the field.

474

Methods

Experiments of mice, birds, and dogs. There are four experiments in this study.

The first is the free-social behavior test of two wild-type mice for the program design of SBeA. 32 adult male C57BL/6 mice (7–12 weeks old) are used for the free-social behavior test. The mice were housed at 4–5 mice per cage under a 12-h light–dark cycle at 22–25 °C with 40–70% humidity, and were allowed to access water and food ad libitum (Shenzhen Institutes of Advanced Technology, Shenzhen, China). Before the social behavior test, the mice are added tail tags using a black mark pen. The tail tags are constructed by horizontal and vertical lines. The horizontal line represents one, and the vertical line represents five. Using the combination of horizontal and vertical lines, the mice are marked according to the sequence of the experiment. After that, the mice are put into a circular open field made of a transparent acrylic wall and white plastic ground, with a base diameter of 50 cm or 20 cm and a height of 50 cm for 5 min or 15 min identity recording one by one using MouseVenue3D. Then, the mice are paired and put into the same circular open field for the free-social behavior test.

The second is the free-social behavior test of mice with different genotypes. 5 adult (8 weeks old) Shank3B knockout (KO; *Shank3B*^{-/-}) mice on C57BL/6J genetic background and 5 adult (8 weeks old) male C57BL/6 mice, were used in the behavioral experiments. *Shank3B*^{-/-} mice were obtained from the Jackson Laboratory (Jax No. 017688) and were described previously²³. The mice were housed at 4–5 mice per cage under a 12-h light–dark cycle at 22–25 °C with 40–70% humidity, and were allowed to access water and food ad libitum (Shenzhen Institutes of Advanced Technology, Shenzhen, China). The mice have added the tail tag introduced above. After that, the mice are put into a circular open field with a base diameter of 20 cm introduced before for 5 min identity recording. Then the mice are paired to WT-WT, WT-KO, and KO-KO groups and put into the same circular open field for the

497 free-social behavior test. The combinations of groups and the sequence of experiments are random
498 generated by customized MATLAB code.

499 The third is the free-social behavior test of two birds. One male and one female *Melopsittacus*
500 *Undulatus* (about 26 weeks old) are used in this experiment. They are housed in a conventional
501 environment with feed regularly (Shenzhen Institutes of Advanced Technology, Shenzhen, China). The
502 birds are first put into a circular open field with a base diameter of 20 cm introduced before for 5 min
503 identity recording one by one and then put in it together for 15 min free-social behavior test and
504 recording.

505 The fourth is the free-social behavior test of two dogs. Two female Belgian Malinois (13 weeks
506 old) are used in this experiment. They are housed in Kunming Police Dog Base of the Chinese Ministry
507 of Public Security, Kunming, 650204, China, and their behavior test of them is finished in the State
508 Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy
509 of Sciences, Kunming, 650223, China. The dogs are first put into a $2 \times 2 \text{ m}^2$ open field made by
510 fences one by one for the identity recording. Restricted by the locomotion of dogs, there are only 6
511 min and 11 min identity frames captured by MouseVenue3D, and both of them are used for
512 identification. Then, they are both put into the open field for 15 min free-social behavior test.

513 All husbandry and experimental procedures of mice and birds in this study were approved by
514 Animal Care and Use Committees at the Shenzhen Institute of Advanced Technology, Chinese
515 Academy of Sciences. And all husbandry and experimental procedures of dogs in this study were
516 approved by Animal Care and Use Committees at the Kunming Institute of Zoology, Chinese Academy
517 of Sciences.

518

519 **MouseVenue3D subtle behavior capture system.** There are three versions of MouseVenue3D
520 systems used in this study.

521 The first version is used for the data capture of the SBM-VIS dataset. Four Intel RealSense D435
522 cameras are mounted orthogonally on four supporting pillars made of stainless steel. The distance
523 between the nearest cameras is 90 cm. The cameras are adjusted to 75 cm off the ground to capture the
524 whole view of the animal activities in the open field. Images were simultaneously recorded at 30
525 frames in 640×480 sizes per second by a PCI-E USB-3.0 data acquisition card and the pyrealsense2
526 Python camera interface package. The cameras are connected to a high-performance computer (i7-
527 9700K, 16G RAM) equipped with a 1-terabyte SSD and 12-terabyte HDD as an image acquisition
528 platform. The computer also controls the camera calibration module.

529 The second version is used for the behavioral capturing of mice and birds. Four Point Grey FLIR
530 Chameleon3 CM3-U3-13S2 cameras with adaptive zoom lenses are mounted orthogonally on four
531 supporting pillars made of stainless steel. The distance between the nearest cameras is 85 cm. The
532 cameras are adjusted to 45 cm off the ground to capture the whole view of the animal activities in the
533 open field. To adapt to the size of the open field, the focal length and the pitch angle of the cameras
534 are flexibly adjusted before each experiment. Images were simultaneously recorded at 30 frames in
535 1288×964 sizes in grayscale per second by a PCI-E USB-3.0 data acquisition card and the Spinnaker
536 Python camera interface package. The cameras are connected to a high-performance computer (i9-
537 10900K, 128G RAM) equipped with a 512-gigabyte SSD and two 16-terabyte HDDs as an image
538 acquisition platform. The computer also controls the camera calibration module.

539 The third version is used for the behavioral capturing of dogs. Four Intel RealSense D435 cameras
540 are mounted orthogonally on walls. The distance between the nearest cameras is 210 cm. The cameras

are adjusted to 150 cm off the ground to capture the whole view of the dog activities in the open field. Images were simultaneously recorded at 30 frames in 640×360 sizes per second by a PCI-E USB-3.0 data acquisition card and the pyrealsense2 Python camera interface package. The cameras are connected to a high-performance computer (i7-9700K, 16G RAM) equipped with a 1-terabyte SSD and 12-terabyte HDD as an image acquisition platform. The computer also controls the camera calibration module.

SBM-VIS Dataset. The free-social behavior of two C57BL/6 mice introduced above is captured by the first version of MouseVenue3D. The first 1 min frames of four cameras are annotated as the SBM-VIS dataset, which is 7200 frames in total. To accelerate the data annotation, we take deep learning for assistance. 30% of the contours are manually labeled, and the rest are firstly labeled by YOLACT++ trained by the manually labeled 30% contours then checked by humans. Then, the single animal DeepLabCut is used to predict the poses of masked frames with the human check. Per 18 frames are grouped for a video instance and saved as YouTubeVIS format⁴⁵. And the poses are saved as a .csv file. The identities across different cameras are corrected by human annotators.

New scenario generation for video instance segmentation. The new scenario generation for video instance segmentation is divided into several steps: contour extraction, trajectory extraction, dataset labeling, background calculation, model self-training, and video dataset generation. After that, it can be input into the instance segmentation model for large-scale training. Suppose the number of animals in the video is n . Conda virtual environment configuration includes OpenCV 4.5.5.62, Python 3.8.12,

Pytorch 1.10.1, The computer was configured with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz and NVIDIA RTX3090 GPU.

In the animal contour step, image thresholding is first done, and then the contour in the image is extracted, and the following formula is used to determine whether the frame is social or not, where i stands for a frame, R_i stands for the judgment result of this frame and num_i stands for the number of contours in this frame:

$$R_i = \begin{cases} social, num_i < n \cap num_i > 0 \\ nonsocial, num_i = n \\ error, others \end{cases} \quad (1)$$

When extracting the animal trajectory, due to the influence of noise, all the contour center points are recorded as the candidates of the animal frame center point, and the closest point to each animal in the previous frame is selected from multiple center points as the true center point of this frame, and then the Hungarian matching idea is used to remove the matching points successfully, to optimize the animal trajectory.

For dataset annotation, different manually annotated datasets were used for different animals. We manually annotated 272 images in the 50 cm mice open field experiment, 805 images in the 20 cm mice open field experiment, 600 images in the birds experiment, and 800 images in the dogs experiment.

For background calculation, the non-mask position (the background) of each image is extracted and fused into the final background image using the labeled data set. The above operation is repeated for all data sets to obtain a clean background image.

The labeled data set is used for YOLACT++ round training, and the trained model is used to predict video frames. The predicted high-quality frames will be added to the original data set for the

583 next round of training. Among them, the selection method of high-quality frames is as follows: i
 584 represents a certain frame, f_i is the segmentation result of the frame i , f_{i-1} is the segmentation result
 585 of the frame $i-1$, F is the calculation process of scoring matrix of all segmentation results in two
 586 frames, the calculation idea refers to the Hungarian matching idea, and the calculation result is G_i :

$$587 \quad G_i = F(f_i, f_{i-1}) \quad (2)$$

588 Then, all G_i are merged and clustered, and the class with the higher overall matrix score is
 589 selected as the high-quality frame class and added to the training data set. YOLACT++ selects the
 590 ResNet50 model as the pre-training model, and the maximum number of iterations is 150,000
 591 generations. The training process takes about 5 hours. After YOLACT++ finishes training, its final
 592 model is used to predict the results for all frames.

593 The video dataset required for instance segmentation training is subsequently generated. The data
 594 set is divided into three parts, which are real data set, social area data set, and randomly generated data
 595 set. The real data set is the continuous high-quality frames predicted and filtered by YOLACT++,
 596 which are written into the video data set after data enhancement, where the data enhancement is
 597 performed by flipping the image left and right. Since there are many occlusions during social
 598 interaction and the performance of the model decreases, it is necessary to generate multiple datasets in
 599 the social area. Here, consecutive frames of animals in the social area are selected and augmented to
 600 generate the social area dataset, where N forms of enhancement are generated by data augmentation,
 601 as shown below, where C represents combination (that is, the combination of different masks is
 602 selected for flipping in each frame). A stands for alignment (that is, all masks are aligned to occlusion):

$$603 \quad N = \left(\sum_n^{i=0} C_n^i \right) * A_n^n \quad (3)$$

604 Since the number of real data and social area data sets may be far from enough to complete the
605 model training task, some data sets in the animal activity area are randomly generated after this step.
606 In this part, the real animal trajectory in the video, the obtained animal mask, and the background
607 calculated in the previous step are used for data collection, and the video data set is written after data
608 enhancement. 14940 video datasets were generated for the 50 cm mice open field experiment, 15130
609 for the 20 cm mice open field experiment, 5970 for the bird experiment, and 41755 for the dog
610 experiment.

611

612 **The training and validation of video instance segmentation model.** Here, the video instance
613 segmentation model adopts the Transformer-based VisTR model, which regards the video instance
614 segmentation task as a parallel sequence encoding and decoding problem. The pre-training model was
615 the ResNet101 model trained on the COCO dataset, the learning rate was set to 0.0001, the dropout
616 parameter was set to 0.1, the training epochs were 30, the frame length was set to 9, the sequence
617 length was set to 19, the number of encoding layers was 6, the number of decoding layers was 6, and
618 Adam was used for the optimizer. The model training takes about 1.5 days. The trained model is
619 evaluated on one minute of standard data, and the model accuracy for video instance segmentation is
620 as follows: IST (Identity swap times) is 5.500 ± 3.640 , ISTP (Identity swap times percentage) is 0.003
621 ± 0.002 , IOU_{NID} (The Intersection of the union without identity) is 0.746 ± 0.017 , mAP50_{NID} (Mean
622 of average precision without identity, the threshold value is greater than 0.5) is 0.985 ± 0.013 , mAP50_{ID}
623 (Mean of average precision with identity, the threshold value is greater than 0.5) is 0.605 ± 0.319 ,
624 similarly, mAP70_{NID} is 0.805 ± 0.068 , mAP70_{ID} is 0.497 ± 0.271 .

625

626 **The single animal poses estimation.** Single animal pose estimation was performed using DeepLabCut
627 2.2.0.4 with a Conda virtual environment with Python 3.8.12. Four different kinds of animals were
628 used in the manual labeling of the dataset, with 709 images labeled for mice in a 50 cm open field,
629 1421 images labeled for mice in a 20 cm open field, 1035 images labeled for birds, and 819 images
630 labeled for dogs. The number of body posture points varied for each animal, with 16 for each mouse
631 (nose, left ear, right ear, neck, left front limb, right front limb, left hind limb, right hind limb, left front
632 claw, right front claw, left hind claw, right hind claw, back, root tail, mid tail, tip tail), 19 for each bird
633 (beak, calvaria, left eye, right eye, neck, left wing root, left wing mid, left wing tip, right wing root,
634 right wing mid, right wing tip, left leg root, left leg tip, right leg root, right leg tip, back, belly, tail root,
635 tail tip), and 17 for each dog (nose, left ear, right ear, neck, left front limb, left front paw, right front
636 limb, right font paw, left hind limb, left hind paw, right hind limb, right hind paw, front back, mid back,
637 hind back, tail root, tail tip). ResNet50 was used as the pre-trained model. The model was trained for
638 a maximum of 103 million iterations with a batch size of 8 and took approximately 10 hours to train
639 on an NVIDIA RTX3090 GPU using Python. The prediction results were saved in a CSV file.

640

641 **3D pose reconstruction of multi-animals.** Here, we use the multi-view geometry method in computer
642 vision for the 3D reconstruction of multiple animals. The basic projection formula between 2D points
643 and 3D space points is as follows.

$$s \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = K \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (4)$$

645 Here, s represents the scaling factor, x and y are the points in the image, K is the camera
646 internal reference, R is the rotation matrix, t is the translation matrix, and X , Y , and Z represent
647 the coordinates of the 3D points. Specifically, firstly, all two-dimensional skeleton information of
648 multi-animal and multi-view was read, and the points in the two-dimensional file with too low a
649 confidence rate were directly set to NaN. Then, the relative position parameters between multiple
650 cameras are read and the triangulation algorithm is used for the 3D reconstruction of a single animal.
651 The basic principle is as follows:

$$\begin{aligned}
&\alpha_1 = K_1[R_1t_1]P \\
&\alpha_2 = K_2[R_2t_2]P \\
&\dots \\
&\alpha_n = K_n[R_nt_n]P
\end{aligned} \tag{5}$$

653 Here, α_1 to α_n represent the two-dimensional points with the same content in different cameras,
654 K_1 to K_n represent the internal parameter matrix of different cameras, R_1 to R_n represent the
655 rotation matrix of different cameras, t_1 to t_n represent the translation matrix of different cameras and
656 the three-dimensional point P can be solved by combining these equations, and we use the SVD
657 decomposition to solve the least squares regression problem.

658 Next, since the appearance of animals in different views is very similar, the identities of instance
659 segmentation may be swapped, and the wrong 3D point coordinates may be calculated. Therefore, we
660 first obtain the full permutation index list of all 2D points of multiple animals in each view angle, and
661 then obtain the 3D point coordinates in all cases. Eventually, the point with the smallest error is selected
662 as the final multi-animal 3D skeleton point.

663

664 **The training and validation of animal identification model.** In multi-animal experiments, because
665 the animal hair is too similar, its identity is likely to be wrongly assigned. Here, we use the deep
666 learning model to identify two-dimensional animals under four perspectives, to determine the identity,
667 and to ensure that the animal identity of the whole video can be corrected back.

668 The data set of identity recognition needs to record the individual activity videos of all experimental
669 animals in the same scene, and then obtain two-dimensional pictures of animals from multiple
670 perspectives. The trained video instance segmentation model is used to predict the mask of the whole
671 body of a single animal (the effect of manually selecting some body parts of animals for identity
672 recognition is not good). Then, the four obtained prediction images are processed by image stretching,
673 stitching, thresholding, and so on, and finally, a complete image is obtained as the training data. The
674 labels are the animal numbers, so there is no need to manually annotate the data. The size of the dataset
675 depends on the duration of recording individual activity videos of animals. In the mouse experiment
676 with a 50 cm open field, the data set size was 594,000, in the mouse experiment with a 20 cm open
677 field, the data set size was 180,000, in the bird social experiment, and the dog social experiment, the
678 data set size was 54161.

679 The deep learning model uses the Efficientnet-b4 model, the maximum number of iterations is set
680 to 120, the initial value of the learning rate is 0.005, and the batch size is set to 32. It is trained on
681 NVIDIA RTX3090 GPU, and each round of training takes about 40 minutes.

682 In the mouse experiment with a 50 cm open field, the accuracy of the identification network in the
683 training set was 0.993, and the accuracy of the validation set was 0.922. In the mouse experiment with
684 a 20 cm open field, the accuracy of the training set was 0.999, and the accuracy of the validation set

was 0.911. In the dog social experiment, the training set accuracy is 0.999, and the validation set accuracy is 0.999.

687

The pattern visualization of animal identification by LayerCAM. LayerCAM can generate the class activation maps (CAM) of each layer of CNN-based models²⁶. The LayerCAM of each layer of the EfficientNet-based identity recognition network is averaged to output a global visualization pattern of animal identities. To further compare the feature weights of different body parts of animals, the 2D poses are used for the body part location of identity frames. From the 2D poses to identity frames, there is a coordinate transformation. The transformed 2D poses on identity frames P_t can be calculated as:

$$P_t = K_r[K_b^1(P_1 + B_b^1), K_b^2(P_2 + B_b^2), \dots, K_b^{cam}(P_{cam} + B_b^{cam})] \quad (6)$$

where K_r is the resized matrix of cascade frames, K_b is the scale matrix of the bounding box of single camera view, P is the raw 2D poses, B_b is the bias matrix of the bounding box of single camera view, and the index cam is the camera number. The K_b is decided by the size of frames and the bounding box size of the cropped animal instance. To reduce the disturbance of 2D pose estimation, a box centered on P_t of each transformed 2D pose crops the LayerCAM value. And the mean value of them represents the CAM weights of each body part.

702

The mask reprojection from 3D poses to video instances. The 3D poses of each animal connect the geometric relationships of the video instances in different camera views. In the step of 3D reconstruction of multi-animals, the 2D poses of each camera view angle have been re-grouped by optimization. Because the 2D poses of multiple animals are constructed by the single animal after

707 video instance segmentation, the masks of instances are matched to the 2D poses. Therefore, the 3D
 708 poses of each animal are corresponding to the masks of video instances frame by frame. A table saves
 709 the corresponding indexes from 3D poses to video instances and is checked frame by frame for mask
 710 reprojection.

711

712 **Parallel decomposition of trajectories.** The parallel decomposition of trajectories includes three parts.

713 The first part is the decomposition of non-locomotor movement. Let X_{ij}^m is the behavior
 714 trajectories of animals m with i frames and j dimensions, the non-locomotor movement component
 715 Y_{NM} can be calculated as follows:

$$716 \quad Y_{NM} = \{X^m - J \frac{1}{N} \sum_{n=1}^N X_{nj}^m\} \quad (7),$$

717 where J is all one vector, and N is the number of frames. After this step, the center of the body
 718 of the animals can be aligned together.

719 The second part is the decomposition of locomotion. The locomotion component Y_L can be
 720 calculated as follows:

$$721 \quad Y_L = \left\{ \frac{\partial X^m}{\partial i} \right\} \quad (8).$$

722 The third part is the decomposition of distance. The distance component Y_D can be calculated as
 723 follows:

$$724 \quad Y_D = \sqrt{(X^1 - X^2)^2} \quad (9).$$

725

726 **Feature representation of distance dynamics.** The distance dynamics Y_{DD} can be calculated as
 727 follows:

$$\begin{cases} Y_{DD} = f_{UMAP}(Y_D), i < I_{thres} \\ Y_{DD} = f_{ResMLP}(Y_D), i \geq I_{thres} \end{cases} \quad (10),$$

where $f_{UMAP}(\cdot)$ is the UMAP mapping including the parameters `n_neighbors` set to 199, and `min_dist` set to 0.3, I_{thres} is the threshold of frames set to 200000, and $f_{ResMLP}(\cdot)$ is the feature representation including ResMLP. For $f_{ResMLP}(\cdot)$, firstly, the Y_D is randomly sampled to Y_{Ds} according to I_{thres} . And the rest of Y_D is Y_{Dr} . Then, Y_{Ds} and $Y_{DDs} = f_{UMAP}(Y_{Ds})$, the UMAP of Y_{Ds} , is used to train ResMLP for feature encoding. After the training, the ResMLP predicts the Y_{DDr} from Y_{Dr} , and the Y_{DD} can be recombined by Y_{DDs} and Y_{DDr} according to the sample point.

The ResMLP is based on the residual module and multi-layer perceptron^{46,47}. The residual block is constructed by multi-layer perceptron with two layers. Each layer has 64 neurons, and two residual blocks are stacked to construct the residual part. The head of ResMLP is one 1d convolution layer and one global max pooling layer for the feature encoding of distance dynamics⁴⁸. The output part of ResMLP is constructed by one fully connected layer with one sigmoid layer for the continuous value representation⁴⁹. The activation function of ResMLP uses ReLU layers⁴⁹. The optimizer of ResMLP is Adam, the initial learning rate is set to 0.001, the mini-batch size is set to 2000, and the epoch number is set to 100⁵⁰. The final RMSE of validation is 0.02~0.06, and the training time of ResMLP is about 4 min on NVIDIA GeForce RTX 3090 GPU.

The time consumption comparison of ResMLP. After the manual time consumption test of UMAP, the quadratic function is used for the estimation time comparison. The coefficient of the quadratic function is 0.00002. The time consumption of ResMLP is estimated as a linear function with a slope

748 set to 0.000008 and an intercept set to 240 based on the training and prediction time of ResMLP. The
 749 functions of the time consumption are as follows:

$$750 \quad T_{UMAP} = k_{UMAP} y_D^2 \quad (11),$$

$$751 \quad T_{ResMLP} = k_{ResMLP} y_D^2 + b_{ResMLP} \quad (12),$$

752 where T_{UMAP} is the time consumption of UMAP, k_{UMAP} is the coefficient of quadratic function, y_D
 753 is the number of distance components, T_{ResMLP} is the time consumption of ResMLP, k_{ResMLP} is the
 754 slope of ResMLP, and b_{ResMLP} is the intercept.

755

756 **The distance map.** Let Y_E is the low-dimensional embedding of the social behavior atlas, and Y_{DM} is
 757 the distance of Y_E . The Y_{DM} can be calculated as follows:

$$758 \quad Y_{DM}^j = \frac{1}{q^j - p^j + 1} \sum_{i=p^j}^{q^j} y_D^i \quad (13),$$

759 where j is one of the points in Y_{DM} , p is the start time point of Y_{DM}^j , and q is the end time
 760 point of Y_{DM}^j .

761

762 **The map/body distance.** The body distance is equivalent to Y_{DM} . The map distance Y_{EM} can be
 763 calculated as follows:

$$764 \quad Y_{EM}^j = \arg \min (J y_E^j - Y_E) \quad (14),$$

765 where y_E is one point of Y_E . And the map/body distance Y_{MB} can be calculated as follows:

$$766 \quad Y_{MB}^j = \frac{Y_{EM}^j}{Y_{DM}^j} \quad (15).$$

767

768 **The adaptive watershed clustering.** The variable of watershed clustering on 2D embeddings is the
 769 kernel bandwidth k_b , which decides the density d . The adaptive watershed clustering is designed to
 770 automatically choose the best d . The best d is determined by the stable number of clusters c_{st} . To
 771 get c_{st} , the clusters under certain k_b are firstly calculated as:

$$772 \quad c_n^i = f_{WC}(Y_E, k_b^i) \quad (16),$$

773 where $f_{WC}(\cdot)$ is the watershed clustering, c_n is the number of clusters. Then, the c_{st} is calculated
 774 as:

$$775 \quad c_{st} = c_n^{f_{Mode}\left(\left|\frac{dc_n^i}{di}\right|\right)} \quad (17),$$

776 where $f_{Mode}(\cdot)$ is the mode function. The c_s is the lower bound of watershed clustering with
 777 larger kernel bandwidth. To improve the sensitivity of watershed clustering for the subtle differences
 778 of social behavior, a threshold u_{thres} is set to 0.9 to restrict k_b in more fine grain. So, the number of
 779 sensitivity clusters c_{se} can be calculated as:

$$780 \quad c_{se} = \arg \max \left(\left| \frac{dc_n^i}{di} \right| - u_{thres} \times \left(f_{Max} \left(\left| \frac{dc_n^i}{di} \right| \right) - f_{Min} \left(\left| \frac{dc_n^i}{di} \right| \right) \right) \right) \quad (18),$$

781 where $f_{Max}(\cdot)$ is the maximum function, and $f_{Min}(\cdot)$ is the minimum function. The c_{st} and c_{se}
 782 together determine the lower and upper bound of watershed clustering.

783
 784 **Behavior mapping of the PAIR-R24M dataset.** The 3D trajectories of the PAIR-R24M dataset are
 785 captured by high-performance cameras with high frame rates. To reduce the processing time and keep
 786 the global features of different mice, the frame rate is downsampled from 120 Hz to 30 Hz. The
 787 classification of the behavioral interactions of the animals includes 4 categories especially close, chase,

788 explore, and NaN value. The NaN value in the social behavior atlas is defined as others. Because the
 789 interaction classes are imbalance in quantity, four coefficients are used to balance the visual effect of
 790 data distribution in the atlas.

791

792 **The cluster purity.** The cluster purity is an indicator to quantify the uniformity of a cluster. Let the
 793 $P = \{p_1, p_2, \dots, p_N\}$ is the ground truth indexes of all data, the $Q = \{q_1, q_2, \dots, q_N\}$ is the cluster indexes
 794 of all data and N is the number of clusters, the cluster purity C_P can be calculated as:

$$795 \quad C_P^i = \frac{\sum p_i \cap q_i}{\sum p_i \cup q_i} \quad (19).$$

796

797 **The cluster gram of grouped mice.** To reveal the inherent patterns of behavior fractions of each group,
 798 the cluster gram is first stacked group by group. Then, all of the behavior fractions are normalized
 799 according to the dimension of the subject and sorted by hierarchical clustering according to the
 800 dimension of the social behavior module. The clustering tree is normalized for better visualization.
 801 Further, the behavior fractions of each group are sorted according to Euclidean distance for the
 802 similarity metric. The initial row of each group for sorting is chosen by the maximum change rate R_m .
 803 The R_m can be calculated as:

$$804 \quad R_m = \sum \left| \frac{ds_m^i}{di} \right| \quad (20),$$

805 where s_m is the sorted social behavior fractions by hierarchical clustering.

806

807 **The angle spectrum clustering.** The angle spectrum clustering is used to merge the similar sub-
808 clusters of behavior in feature vector space. Let V is the feature vector matrix of social behavior
809 modules in PCA space, the angle spectrum A_s can be calculated as:

$$810 \quad a_s^{ij} = \arccos \left(\frac{v_i \cdot v_j}{|v_i| \times |v_j|} \right) \quad (21),$$

811 where v is one of the feature vectors in V . Then, the A_s is clustered by hierarchical clustering
812 according to the 11 components of 99% variance explanation.

813

814 **Computational software and hardware.** The development of the 3D tracking part of SBeA is based
815 on Python 3.8.12 in the Conda environment on Ubuntu 20.04. The development of the behavior
816 mapping part and figure plot uses MATLAB R2021b. All of the statistics are finished by Prism 8.0
817 (GraphPad Software). The development of SBeA is on a high-performance workstation with two Intel
818 Xeon Silver 4210R, eight NVIDIA GeForce RTX 3090, 2 Tera Byte RAM, and a 140 Tera Byte
819 Network Attached Storage. SBeA has been verified to be able to be applied in a workstation with one
820 Intel i9-12900K CPU, at least one NVIDIA GeForce RTX 3090 GPU, and 128 Giga Byte RAM.

821

822 **Statistics.** Before hypothesis testing, data were first tested for normality by the Shapiro–Wilk
823 normality test and for homoscedasticity by the F test. For normally distributed data with homogeneous
824 variances, parametric tests were used; otherwise, non-parametric tests were used. All of the ANOVA
825 analyses are corrected by the recommended options of Prism 8.0. No data in this work are removed.
826 All related data are included in the analysis.

827

828 **Data availability**

829 We provide all model weights and the validation data used in this paper at
830 https://figshare.com/projects/Social_behavior_atlas/162718 . For this study, we collected more than 1
831 TB of video data for animal social behavior, and they were processed by SBeA step by step. We also
832 provide necessary intermediate data for the validation and visualization of this study. This paper can
833 be replicated by the downloadable data and supplied code.

834

835

836 **Code availability**

837 We provide a code repository of SBeA at https://github.com/YNCRis/SBeA_release . This repository
838 includes SBeA_tracker and SBeA_mapper. SBeA_tracker achieves 3D pose tracking, which has a
839 software interface to guide its usage. SBeA_mapper achieves the atlas mapping of social behaviors
840 from 3D pose trajectories with different configurations. It also contains the code to replicate the figures
841 of this paper.

842

843

References

1. Stanley, D. A. & Adolphs, R. Toward a neural basis for social behavior. *Neuron* vol. 80 Preprint at <https://doi.org/10.1016/j.neuron.2013.10.038> (2013).
2. Agezo, S. & Berman, G. J. Tracking together: estimating social poses. *Nature Methods* 2022 19:4 **19**, 410–411 (2022).
3. Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods* 2022 19:4 **19**, 496–504 (2022).
4. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods* 2022 19:4 **19**, 486–495 (2022).
5. Marks, M. *et al.* Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature Machine Intelligence* 2022 4:4 **4**, 331–340 (2022).
6. 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. *Nature Methods* 2019 16:2 **16**, 179–182 (2019).
7. Ro, S. *et al.* Simple Behavioral Analysis (SimBA) – an open source toolkit for computer classification of complex social behaviors in experimental animals. *bioRxiv* 2020.04.19.049452 (2020) doi:10.1101/2020.04.19.049452.
8. Chen, Z. *et al.* AlphaTracker: A Multi-Animal Tracking and Behavioral Analysis Tool. *bioRxiv* 2020.12.04.405159 (2020) doi:10.1101/2020.12.04.405159.
9. Han, Y., Huang, K., Chen, K., Wang, L. & Wei, P. An automatic three dimensional markerless behavioral tracking system of free-moving mice. *2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems, CYBER 2021* 306–310 (2021) doi:10.1109/CYBER53097.2021.9588299.
10. Huang, K. *et al.* A hierarchical 3D-motion learning framework for animal spontaneous behavior mapping. *Nat Commun* **12**, (2021).
11. Vidal, M., Wolf, N., Rosenberg, B., Harris, B. P. & Mathis, A. Perspectives on Individual Animal Identification from Biology and Computer Vision. *Integr Comp Biol* **61**, 900–916 (2021).
12. Marshall, J. D. *et al.* Continuous Whole-Body 3D Kinematic Recordings across the Rodent Behavioral Repertoire. *Neuron* **109**, 420–437.e8 (2021).
13. Ghiasi, G. *et al.* Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. 2918–2928 Preprint at <https://cocodataset.org/> (2021).
14. Xu, Z. *et al.* Continuous Copy-Paste for One-Stage Multi-Object Tracking and Segmentation. 15323–15332 Preprint at <http://www.cvlb.net/> (2021).
15. Weiss, K., Khoshgoftaar, T. M. & Wang, D. D. A survey of transfer learning. *J Big Data* **3**, 1–40 (2016).
16. Zhuang, F. *et al.* A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE* vol. 109 Preprint at <https://doi.org/10.1109/JPROC.2020.3004555> (2021).
17. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).
18. Liu, N. *et al.* Objective and comprehensive re-evaluation of anxiety-like behaviors in mice using the Behavior Atlas. *Biochem Biophys Res Commun* **559**, 1–7 (2021).
19. Han, Y. *et al.* MouseVenue3D: A Markerless Three-Dimension Behavioral Tracking System for Matching Two-Photon Brain Imaging in Free-Moving Mice. *Neurosci Bull* **38**, 303–317 (2022).

- 887 20. Bolya, D., Zhou, C., Xiao, F. & Lee, Y. J. YOLACT++ Better Real-Time Instance Segmentation. *IEEE Trans*
888 *Pattern Anal Mach Intell* **44**, 1108–1121 (2022).
- 889 21. Bolya, D., Fanyi, C. Z., Yong, X. & Lee, J. YOLACT Real-time Instance Segmentation. *openaccess.thecvf.com*
890 <https://github.com/dbolya/yolact>. (2019).
- 891 22. Wang, Y. *et al.* End-to-End Video Instance Segmentation With Transformers. 8741–8750 Preprint at
892 <https://git.io/VisTR> (2021).
- 893 23. Peça, J. *et al.* Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature* **472**, 437–
894 442 (2011).
- 895 24. Mei, Y. *et al.* Adult restoration of Shank3 expression rescues selective autistic-like phenotypes. *Nature* **530**,
896 481–484 (2016).
- 897 25. Tan, M. & Le, Q. v. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 6105–6114
898 Preprint at <https://proceedings.mlr.press/v97/tan19a.html> (2019).
- 899 26. Jiang, P. T., Zhang, C. bin, Hou, Q., Cheng, M. M. & Wei, Y. LayerCAM: Exploring hierarchical class
900 activation maps for localization. *IEEE Transactions on Image Processing* **30**, 5875–5888 (2021).
- 901 27. Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S. & de Polavieja, G. G. IdTracker: Tracking
902 individuals in a group by automatic identification of unmarked animals. *Nat Methods* **11**, 743–748 (2014).
- 903 28. Ebbesen, C. L. & Froemke, R. C. Body language signals for rodent social communication. *Curr Opin*
904 *Neurobiol* **68**, 91–106 (2021).
- 905 29. Bzdok, D. & Dunbar, R. I. M. The Neurobiology of Social Distance. *Trends Cogn Sci* **24**, 717–733 (2020).
- 906 30. von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep
907 behavior profiling. *Neuropsychopharmacology* 1–12 (2020) doi:10.1038/s41386-020-0751-7.
- 908 31. Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M. & Mainen, Z. F. Big behavioral data: Psychology,
909 ethology and the foundations of neuroscience. *Nature Neuroscience* Preprint at
910 <https://doi.org/10.1038/nn.3812> (2014).
- 911 32. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension
912 Reduction. (2018).
- 913 33. Shi, S., Wang, Y., Dong, H., Gui, G. & Ohtsuki, T. Smartphone-Aided Human Activity Recognition Method
914 using Residual Multi-Layer Perceptron. *INFOCOM WKSHPs 2022 - IEEE Conference on Computer*
915 *Communications Workshops* (2022) doi:10.1109/INFOCOMWKSHPs54753.2022.9798274.
- 916 34. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–1135 (2015).
- 917 35. Wu, Y. E. *et al.* Neural control of affiliative touch in prosocial interaction. *Nature* 2021 599:7884 **599**, 262–
918 267 (2021).
- 919 36. Marshall, J. D. *et al.* The PAIR-R24M Dataset for Multi-animal 3D Pose Estimation. *bioRxiv*
920 2021.11.23.469743 (2021) doi:10.1101/2021.11.23.469743.
- 921 37. Day, F. R., Ong, K. K. & Perry, J. R. B. Elucidating the genetic basis of social interaction and isolation. *Nat*
922 *Commun* **9**, (2018).
- 923 38. Wu, Y. E. & Hong, W. Neural basis of prosocial behavior. *Trends Neurosci* (2022)
924 doi:10.1016/J.TINS.2022.06.008.
- 925 39. Dunn, T. W. *et al.* Geometric deep learning enables 3D kinematic profiling across species and environments.
926 *Nature Methods* 2021 18:5 **18**, 564–573 (2021).
- 927 40. Mathis, A. *et al.* Pretraining boosts out-of-domain robustness for pose estimation. in *Proceedings - 2021 IEEE*
928 *Winter Conference on Applications of Computer Vision, WACV 2021* (2021).
929 doi:10.1109/WACV48630.2021.00190.
- 930 41. Li, W. *et al.* Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation. *IEEE*

Trans Multimedia (2022) doi:10.1109/TMM.2022.3141231.

42. Vaswani, A. *et al.* *Attention Is All You Need*. (2017).
43. Sun, J. J. *et al.* The Multi-Agent Behavior Dataset: Mouse Dyadic Social Interactions. (2021).
44. 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**, 1–73 (2021).
45. Yang, L., Fan, Y. & Xu, N. Video instance segmentation. in *Proceedings of the IEEE International Conference on Computer Vision* vols 2019-October (2019).
46. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the {IEEE} conference on computer vision and pattern recognition* 770–778 (2016). doi:10.1109/CVPR.2016.90.
47. Kruse, R., Mostaghim, S., Borgelt, C., Braune, C. & Steinbrecher, M. Multi-layer Perceptrons. 53–124 (2022) doi:10.1007/978-3-030-42227-1_5.
48. Kiranyaz, S. *et al.* 1D convolutional neural networks and applications: A survey. *Mech Syst Signal Process* **151**, (2021).
49. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* Preprint at <https://doi.org/10.1038/nature14539> (2015).
50. Zhang, Z. Improved Adam Optimizer for Deep Neural Networks. in *2018 IEEE/ACM 26th International Symposium on Quality of Service, IWQoS 2018* (2019). doi:10.1109/IWQoS.2018.8624183.
51. Kort, R. *et al.* Shaping the oral microbiota through intimate kissing. *Microbiome* **2**, (2014).
52. Clucas, B. Patterns of Behavior: Konrad Lorenz, Niko Tinbergen, and the Founding of Ethology. *J Mammal* **87**, (2006).
53. Kaminski, J. & Marshall-Pescini, S. *The Social Dog: Behavior and Cognition*. *The Social Dog: Behavior and Cognition* (2014). doi:10.1016/C2012-0-06593-3.
54. de Chaumont, F. *et al.* Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning. *Nature Biomedical Engineering* 2019 3:11 **3**, 930–942 (2019).

957 **Acknowledgements**

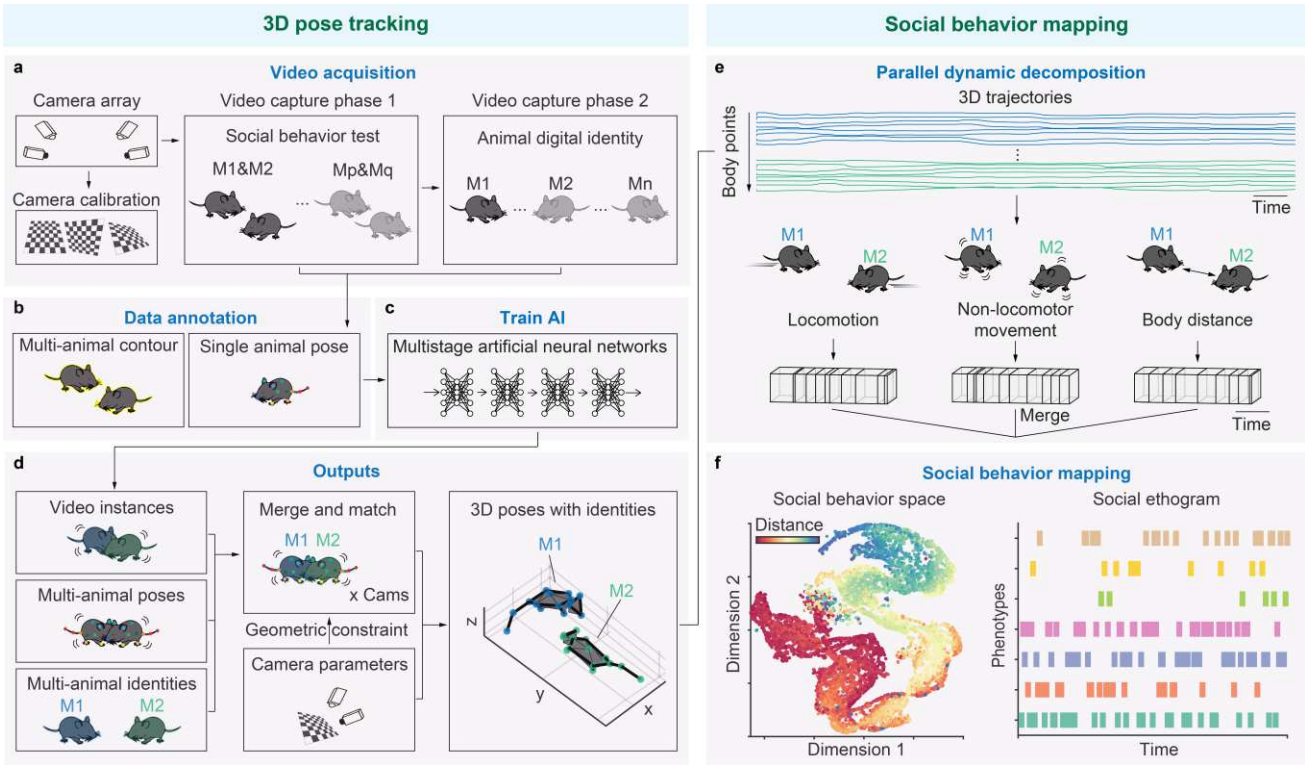
958 This work was supported in part by STI2030-Major Projects (2021ZD0203900), National Natural
959 Science Foundation of China (32222036), the Youth Innovation Promotion Association of the Chinese
960 Academy of Sciences (Y2021100), the National Key R&D Program of China (2018YFA0701403),
961 CAS Key Laboratory of Brain Connectome and Manipulation (2019DP173024), and Guangdong
962 Provincial Key Laboratory of Brain Connectome and Behavior (2017B030301017). We thank
963 ChatGPT for the English language editing of this paper. We thank Jiajia Zhang, Yongji Wu, and Gao
964 Gao for the preliminary tests and experiments. We thank Zhiwei Jiang for comments on our manuscript.

965 **Author contributions**
966 Conceptualization was done by YN. H., K. C. and PF. W. Code was done by YN. H., K. C., and ZW.
967 W. User interface was done by ZW. W. Algorithm design was done by YN. H. and K. C. Mouse data
968 were gathered by YN. H., WH. L., XJ. W., and YT. H. Bird data were gathered by YN. H., CL. H.,
969 and YT. H. Dog data were gathered by JX. L., YWZ. S., N. W., J. L., GD. W., YP. Z., YN. H., YT.
970 H., XJ. W., and JH. L. Hardware was set up by YN. H. and K. H. Data analysis was done by YN. H.
971 Experiments were supported by LP. W. and PF. W. The article was written by YN. H., K. C., YK. W.,
972 and PF. W. with input from all authors. PF. W. supervised the project.
973

974 **Competing interests**

975 The authors declare no competing interests.

976



978 **Fig. 1| The architecture of Social Behavior Atlas.** **a**, Video acquisition for free social behavior test. The camera array is used for behavioral capturing, and it is calibrated by checkboard images. There are two phase for beahvioral video capturing including social behavior test and animal digital identity. The phase 1 is to capture the videos of free-social interactions of two mice. The phase 2 is to capture the identities of each mice in phase 1. **b**, Data annotation for AI training. Social Behavior Atlas need the annotations of multi-animal contour and single animal pose. **c**, The multistage artificial neural networks for 3D pose tracking. **d**, The outputs of 3D pose tracking. Left: The outputs of AI including video instances, multi-animal poses, and multi-animal identities. Center: Combining video instances, multi-animal poses, and multi-animal identities with camera calibration parameters for 3D reconstruction with identities. Right: The visualization of 3D poses with identities. **e**, Parallel dynamic decomposition of body trajectories. Raw 3D trajectories of two animals can be decomposed into locomotion, non-locomotor movement and body distance. After dynamical temporal decomposition, these three parts are merged together as social behavior motifs for behavioral mapping. **f**, Social behavior metric. Social behavior motifs are clustered and pheontyped according to the distribution in social behavior space.

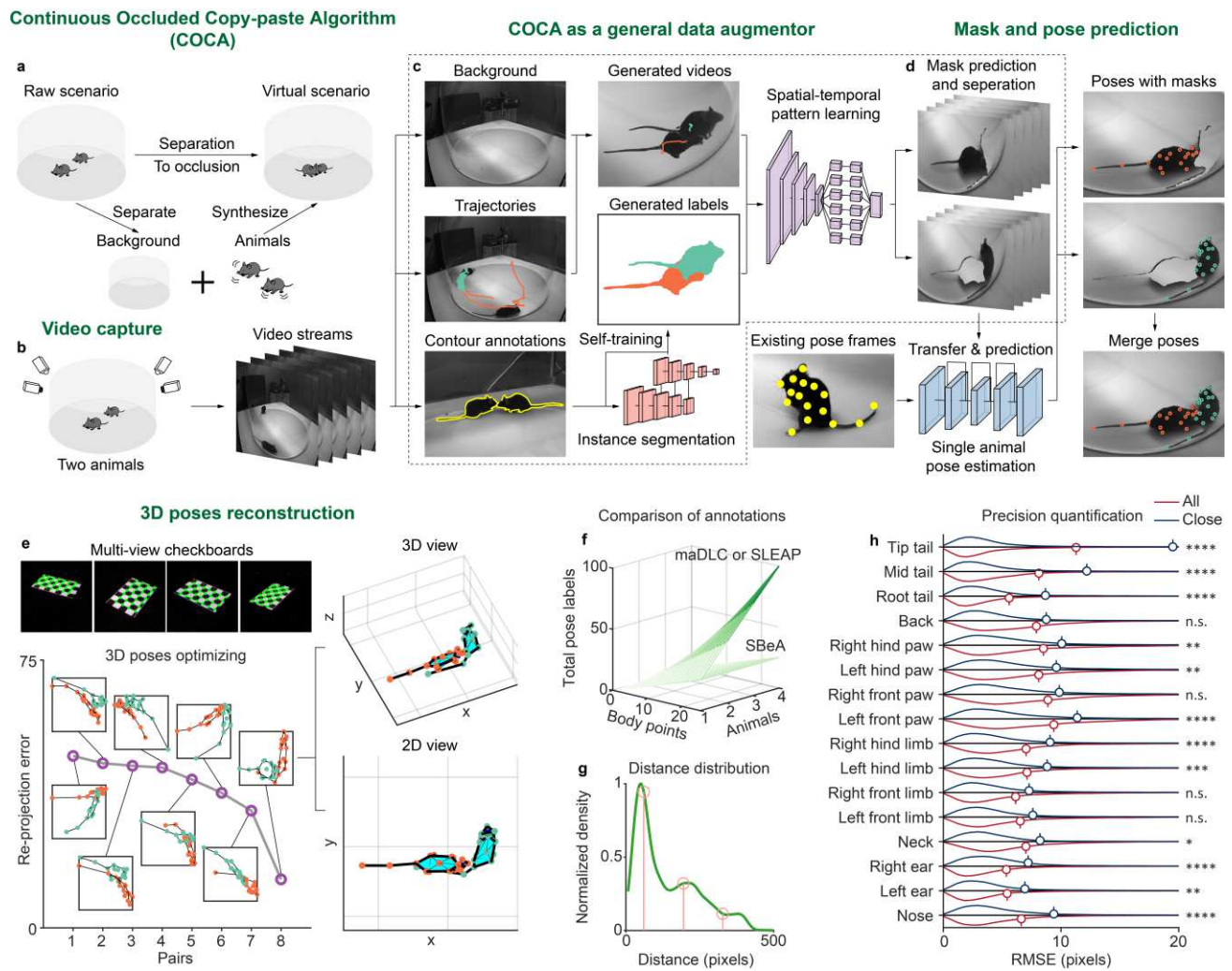


Fig. 2| Continuously occluded copy-paste algorithm (COCA) as a data augmentor for multi-animal tracking. **a**, Concept diagram of COCA. From raw scenario, the instances of background and animals can be synthesized with occlusion in new combination. That achieves generating big data from small data. **b**, Video capture of two free-moving animals. Two animals are put in transparent circular open field and the video streams of behavior are captured by camera array. **c**, COCA as a general augmentor for multi-animal patching according to a little manually labeled data. Behavioral video streams are separated to backgrounds (top left), trajectories (medium left) and manually labeled masks (bottom left). Self-training instance segmentation model is used to predict more unlabeled masks from manually labeled masks. They then combined with backgrounds and trajectories to generate new scenarios of two free-moving mice. **d**, Mask and pose prediction. Spatial-temporal learning is used for the new scenarios and predict the masks of real mouse instances. Then, the single animal pose estimation model can be used for each animal and further the 2D pose of them are merged together to achieve multi-animal pose estimation. **e**, 3D poses reconstruction. The camera array are calibrated by checkboard images using Zhang's calibration. And reprojection errors of all combination pairs of 2D poses of each animals are optimized for 3D reconstruction. Top right: 3D view of 3D poses of two mice in this case. Bottom right: 2D view of 3D poses of two mice in this case. **f**, Comparison of the number of manually labeled points of SBeA and maDLC. **g**, Distance distribution of two free-moving mice. Pink stems are distance boundaries clustered by k-means (close: 60.69, interim: 195.03, far: 327.47). **h**, Prediction error comparison of all validation data. The differences between all and close

1016 data are about ± 2 pixels (two-way ANOVA followed by Sidak multiple comparisons test). RMSE:
1017 root-mean squared error, n.s.: no significant difference, *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$, ****:
1018 $P < 0.0001$.
1019
1020

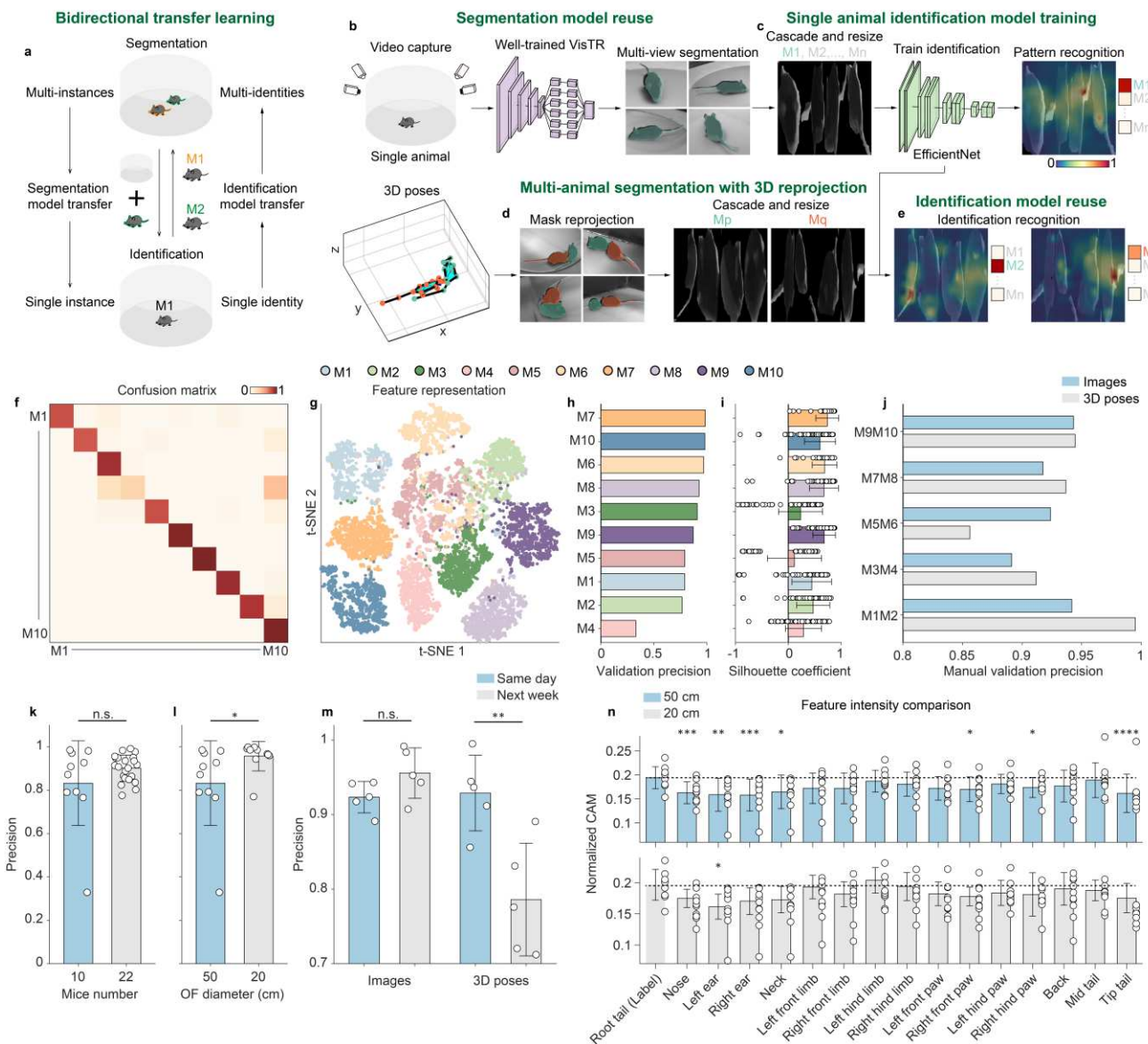


Fig. 3| Bidirectional transfer learning-based animal identification. **a**, Concept diagram of bidirectional transfer learning-based animal identification. Well trained segmentation model on multi-animals can be transferred to the single animal, and well trained identity recognition model on the single animal can also be transferred to multi-animals. The transfer learning of two models reduces unnecessary manual annotations of animal identities. **b**, Segmentation model reuse. Left: an animal is put in transparent circular open field and the video streams are captured by camera array. Center: The well-trained VisTR is reused for the single animal. Right: The output of well-trained VisTR on the single animal. **c**, Single animal identification model training. Left: the single animal instances of multi-view are cropped, cascaded and resized to an image. Center: using EfficientNet as the backbone to train multi-animal classifier. Right: The identity recognition pattern visualization by LayerCAM. **d**, Multi-animal segmentation with 3D reprojection. Left: mask reprojection of each camera view. Right: crop, cascade and resize of two animal instances from matched camera view angles. **e**, Identification model reuse. The well-trained identification model on the single animal can be reused in multi-animal identification. **f**, Confusion matrix of single animal identification. **g**, Feature representation of single animal identification using t-SNE. **h**, The sorted validation precision of **f**. **i**, The sorted silhouette

1037 coefficient of **g**. **j**, The manual validation precision of multi-animal identification. **k**, The identification
1038 precision under different mice number. The identification of 10 mice uses 7200 frames for training and
1039 1800 for validation, and 22 mice uses 21600 frames for training and 5400 frames for validation. With
1040 the increase of animal number, the add of training frames can keep higher identification precision (two-
1041 sided Mann–Whitney test). **l**, The identification precision under different open field (OF) diameter of
1042 behavioral test(two-sided Mann–Whitney test). **m**, The identification precision in different interval
1043 time between social behavior test and identify recording(two-sided unpaired T-test). **n**, The feature
1044 intensity of the tracking body parts under different OF diameter of behavioral test. The root tail of mice
1045 is labeled by different black line markers for the easy distinguish of human(one-way ANOVA followed
1046 by Dunnett multiple comparisons test). n.s.: no significant difference, *: P<0.05, **: P<0.01, ***:
1047 P<0.001, ****: P<0.0001.
1048
1049
1050

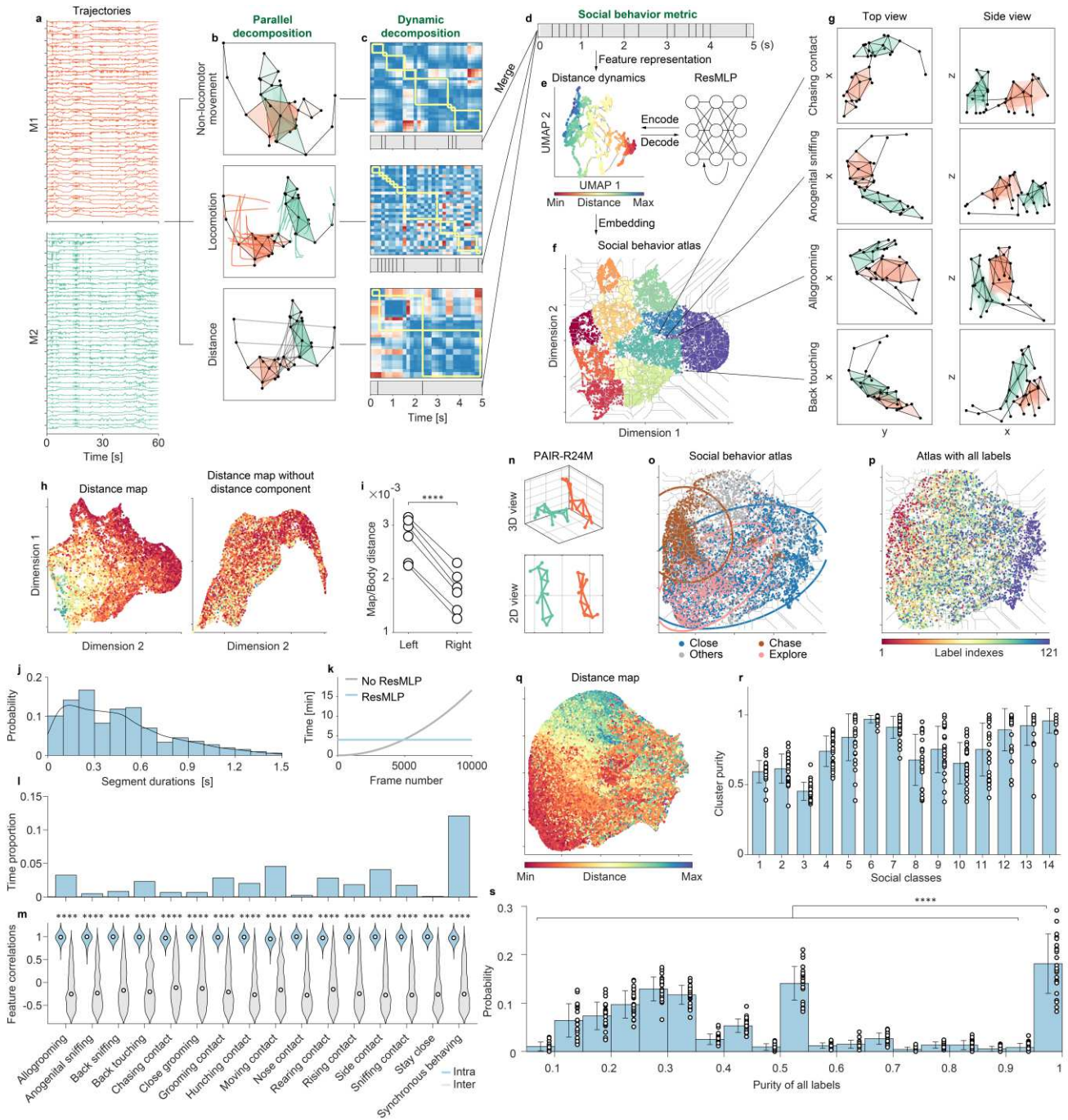


Fig. 4| Natural behavioral structure-inspired segmentation and mapping of free social interactions. **a**, The 3D trajectories of 2 animals. **b**, The parallel decomposition of trajectories. Top: Non-locomotor movement. Middle: Locomotion. Bottom: Distance. **c**, The dynamic decomposition after parallel decomposition using Dynamic Time Alignment Kernel (DTAK). **d**, **e**, and **f**, Social behavior metric after dynamic decomposition. **d**, Decomposed segments merging. **e**, Feature representation of segments. Left: Dimensional reduction of distance dynamics. Right: Residual Multi-Layer Perceptron (ResMLP) for feature refining. **f**, Social behavior atlas construction. The adaptive watershed is used for clustering. Color represents large clusters and area represent sub-clusters. **g**, Social behavior cases clustered in social behavior atlas. **h**, Left: The distance map of **f**, Right: The distance map without the distance component in behavior decomposition steps. **i**, The comparison of map/body distance of **h** (two-sided paired T-test). The higher the map/body distance, the better

1063 performance for the representation of social behavior with different distances. **j**, The probability of
1064 segment durations. **k**, The comparison of computational time consumption of feature representation
1065 with or without ResMLP. **l**, The time proportion of different behavior. **m**, The feature correlations intra
1066 and inter behavioral classes (two-sided Mann–Whitney test). **n-s**, The performance quantification of
1067 SBeA on the PAIR-R24M dataset. **n**, The visualization of two mice in the PAIR-R24M dataset. **o**, The
1068 social behavior atlas of PAIR-R24M dataset. The social classes of the PAIR-R24M dataset are
1069 separated in social behavior atlas. The ellipse is the Gaussian model fitting of the three classes. **p**, The
1070 social behavior atlas of all the class labels of PAIR-R24M dataset. The 11 classes of each mouse are
1071 combined to 121 classes, and the 121 classes are distributed with patterns. **q**, The distance map of
1072 social behavior atlas. The distance distribution of distance map is coincident with labels in **o**. **r**, The
1073 cluster purity of social classes in **o**. **s**, The cluster purity probability of all labels in **p**. The cluster
1074 purities greater than 0.95 are significant higher than others (one-way ANOVA followed by Tukey
1075 multiple comparisons test). n.s.: no significant difference, *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$, ****:
1076 $P < 0.0001$.
1077
1078

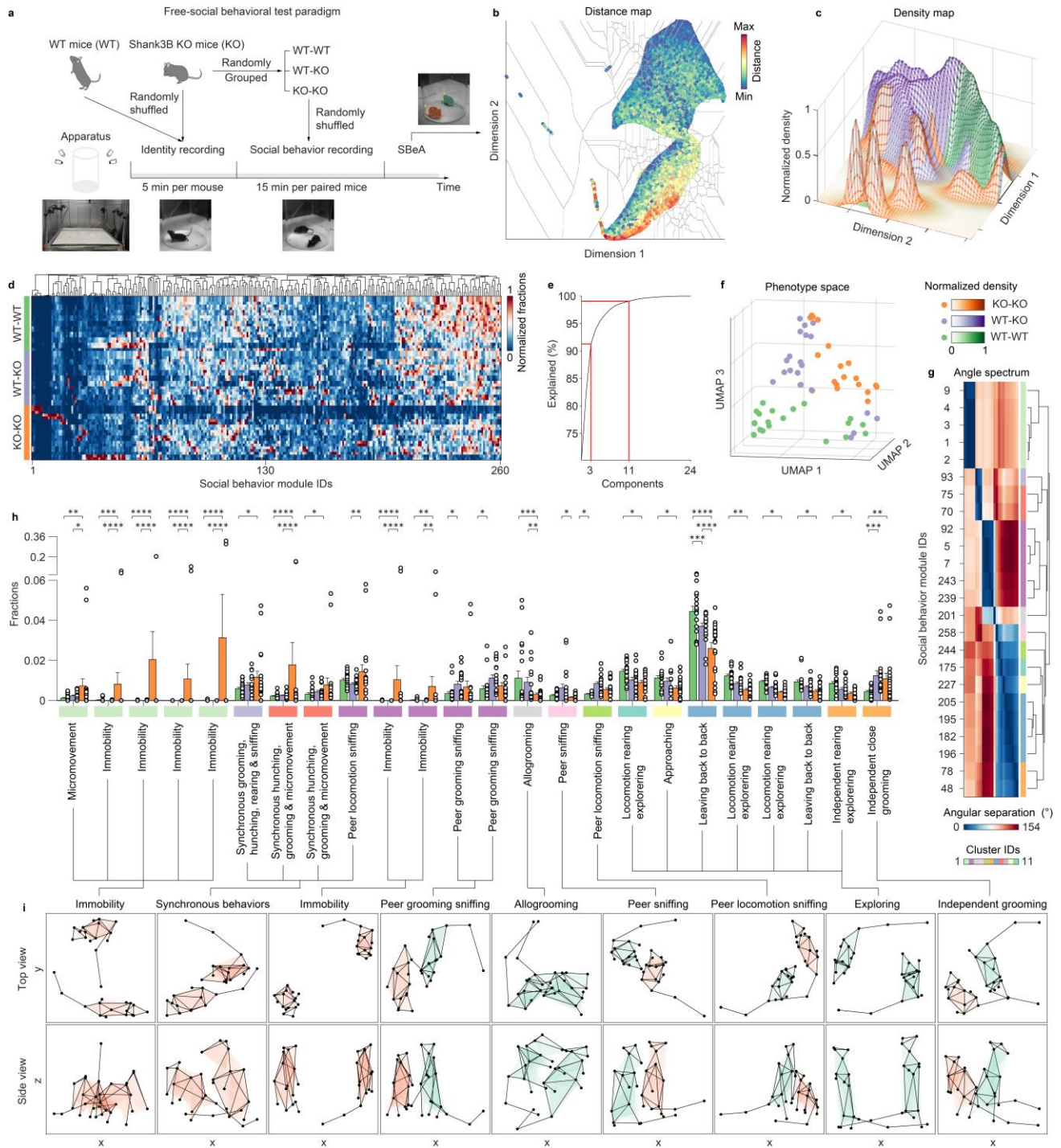


Fig. 5| The identifying of abnormal social behavior modules in *Shank3B* knockout mice. a, The paradigm of free-social behavioral test. WT: Wild type mice, KO: *Shank3B* knockout mice. **b**, The social behavior atlas with distance map of 3 grouped mice. **c**, The distribution of social behavioral modules of three social groups. A total of 260 social behavior modules are identified. **d**, The fractions of social behavioral modules of three social groups. The fractions of each group are normalized, and they are clustered and resorted according to the dimension of social behavior modules. **e**, Dimensional reduction of behavior fractions using principal component analysis (PCA) after hypothesis testing (two-way ANOVA followed by Tukey multiple comparisons test). 24 social behavior modules are significant differences in three groups. 3 components can explain more than 90% variances, and 11

components can explain more than 99% variances. **f**, The construction of phenotype space. UMAP is used to reduce the 260 dimensions of social behavior modules to 3 dimensions according to **e**. Different colors of dots represent different social groups. The phenotypes of 3 social groups can be separated in phenotype space. **g**, The merging of social behavior modules according to behavioral feature angles and **e**. 24 social behavior modules with significant differences are firstly mapped to PCA feature space and then the angular separation are calculated to construct angle spectrum. Further, hierarchical clustering is used to cluster angle spectrum to 11 clusters according to **e**. **h**, The comparison of behavioral fractions of 3 social groups. 24 social behavior modules with significant differences are manually identified. **i**, The visualization of merged social behavior modules. With the assistance of **g**, 9 social behavior modules are merged and identified from 24 social behavior modules. The color of mice represented the behavior cases with the highest mean fraction in **h**. Orange 3D mice represent KO mice, and green 3D mice represent WT mice. n.s.: no significant difference, *: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$, ****: $P < 0.0001$.

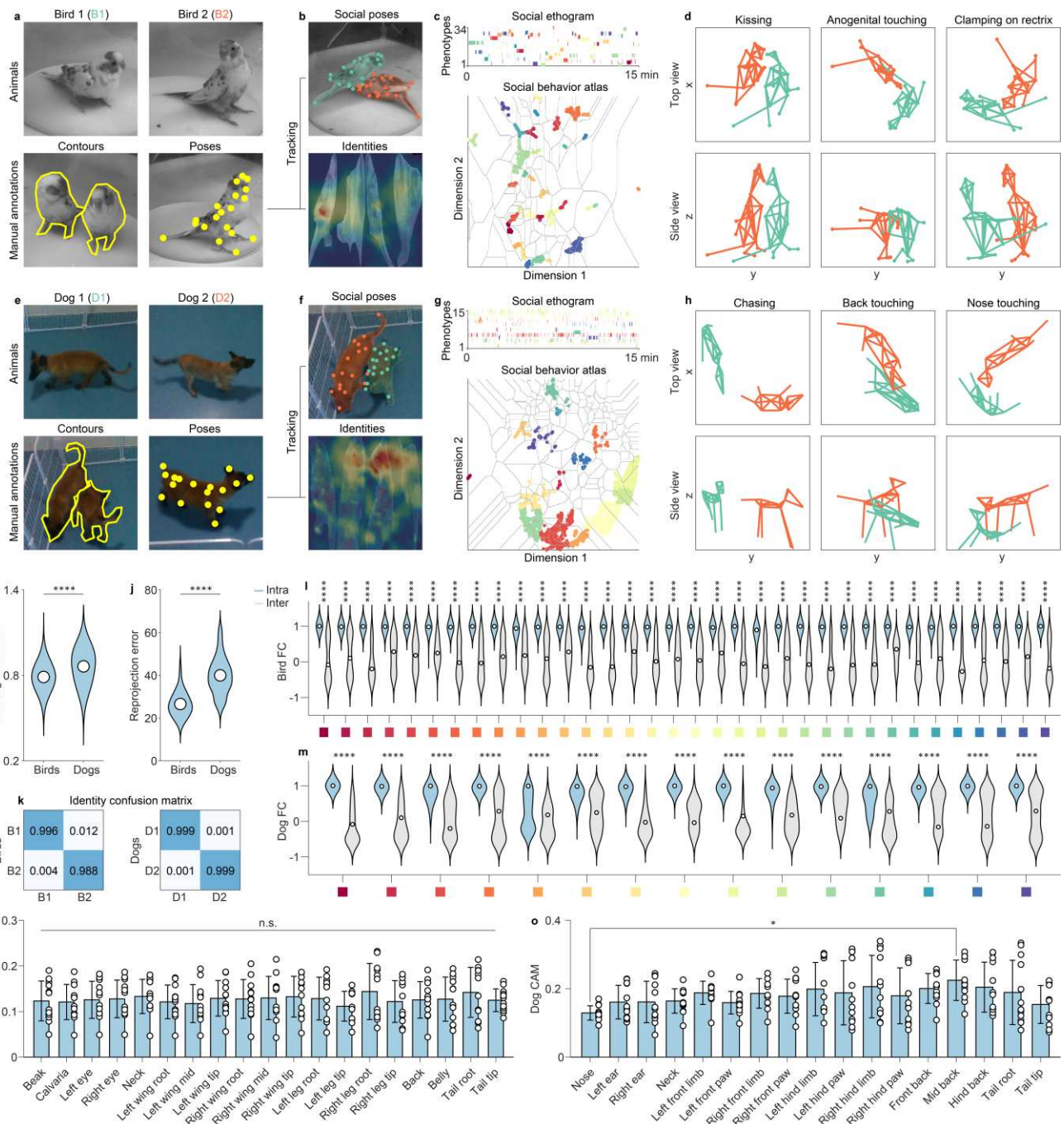
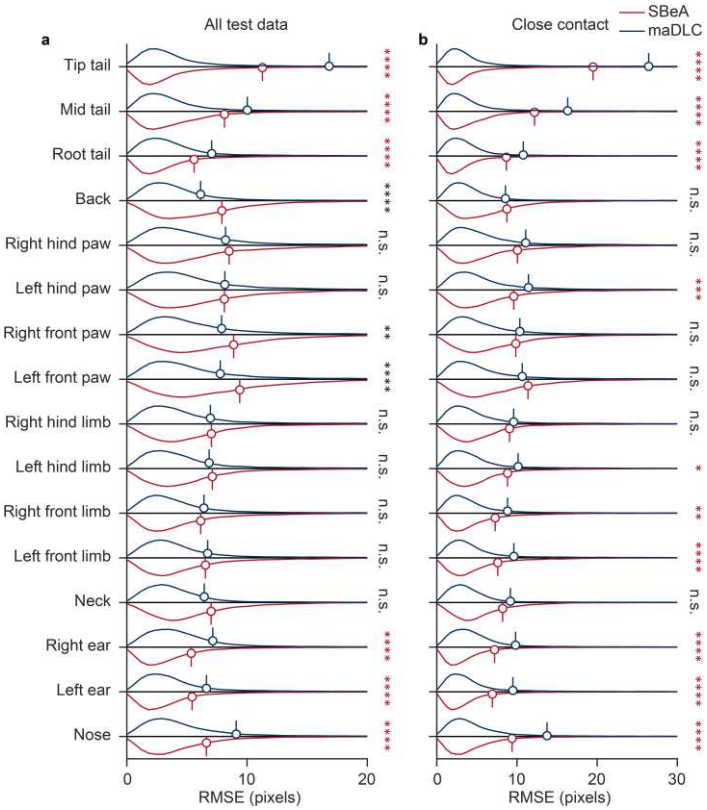


Fig. 6| SBeA for the applications across species such as birds and dogs. a-d, SBeA is used for birds. **a**, The preparation of birds. Two parrots with inconspicuous appearance difference are used for social behavior test. After video recording of identity and free-social behavior by camera array, the contours and poses are manually annotated. 19 body parts are defined for 3D pose tracking. **b**, The social poses and identities outputs of SBeA. **c**, The social ethogram and social behavior atlas of birds. **d**, The 3D social behavior cases of birds. **e-h**, SBeA is used for dogs. **e**, The preparation of dogs. Two Belgian Malinois with inconspicuous appearance difference are used for the social behavior test. After video recording of identity and free-social behavior by camera array, the contours and poses are manually annotated. 17 body parts are defined for 3D pose tracking. **f**, The social poses and identities outputs of SBeA. **g**, The social ethogram and social behavior atlas of dogs. **h**, The 3D social behavior cases of dogs. **i-o**, The performance quantification of SBeA in birds and dogs. **i**, The tracking likelihoods of

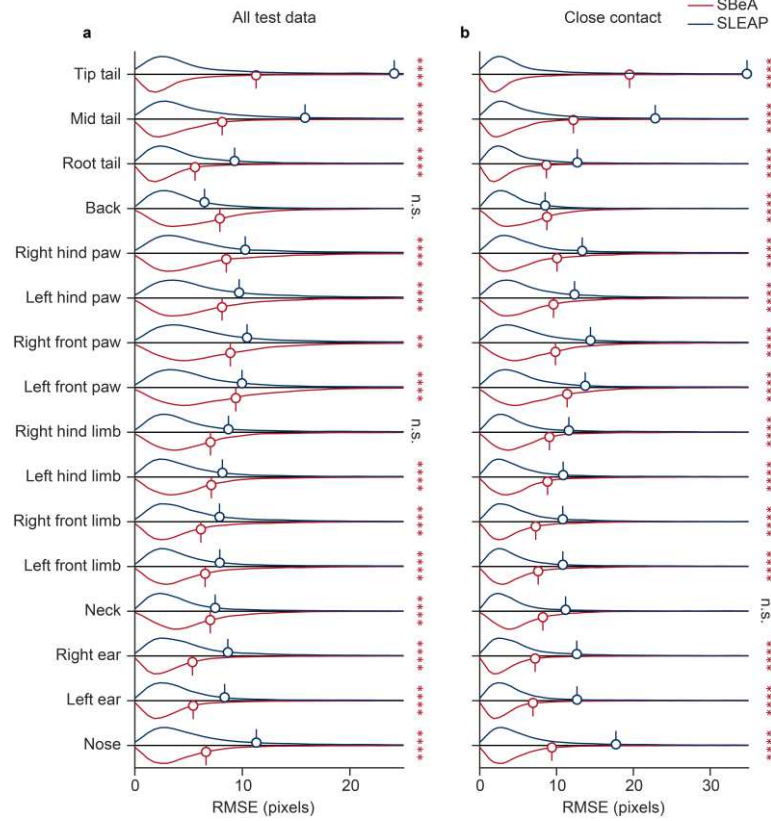
1117 birds and dogs are significant different (two-sided Mann–Whitney test). **j**, The 3D reprojection error
1118 of birds and dogs are significant different (two-sided Mann–Whitney test). **k**, The identity recognition
1119 confusion matrix of birds and dogs. **l**, The feature correlations (FC) intra and inter behavioral classes
1120 of birds (two-way ANOVA followed by Sidak multiple comparisons test). **m**, The FC intra and inter
1121 behavioral classes of dogs (two-way ANOVA followed by Sidak multiple comparisons test). **n**, The
1122 feature intensity of the tracking body parts of birds (one-way ANOVA followed by Tukey multiple
1123 comparisons test).The feature intensities do not show significant differences. **o**, The feature intensity
1124 of the tracking body parts of dogs (one-way ANOVA followed by Dunnett multiple comparisons
1125 test).The feature intensities between nose and mid back show significant differences. n.s.: no
1126 significant difference, *: $P<0.05$, **: $P<0.01$, ***: $P<0.001$, ****: $P<0.0001$.
1127



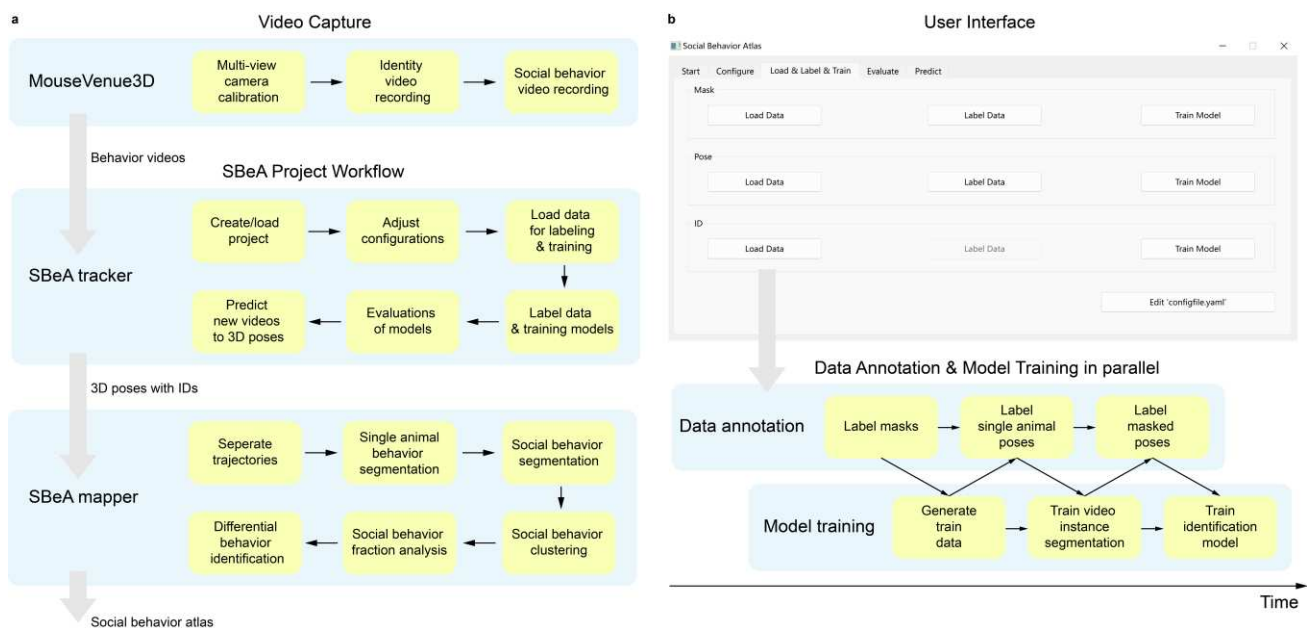
1129

1130 **Extended Data Fig. 1| Performance comparison of SBeA and maDLC. a**, Prediction error
1131 compasion of all test data. The RMSE of most of the body parts of SBeA is significantly lower than
1132 maDLC (two-way ANOVA followed by Sidak multiple comparisons test). **b**, Prediction error
1133 compasion of close contact. The RMSE of all of the body parts of SBeA is significantly lower than
1134 maDLC or even with maDLC (two-way ANOVA followed by Sidak multiple comparisons test). RMSE:
1135 root-mean squared error, n.s.: no significant difference, *: P<0.05, **: P<0.01, ***: P<0.001, ****:
1136 P<0.0001.

1137



Extended Data Fig. 2| Performance comparison of SBeA and SLEAP. a, Prediction error compasion of all test data (two-way ANOVA followed by Sidak multiple comparisons test). **b,** Prediction error compasion of close contact (two-way ANOVA followed by Sidak multiple comparisons test). RMSE: root-mean squared error, n.s.: no significant difference, *: $P<0.05$, **: $P<0.01$, ***: $P<0.001$, ****: $P<0.0001$.



Extended Data Fig. 3| SBeA project workflow. **a**, The workflow from video capture to data processing. **b**, The user interface of SBeA. In order to save the time from video capture to social behavior atlas, the user interface is designed in multithreading. It achieves data annotation and model training in parallel.

Social behavior	Definition	Species
Approaching	One individual approaching another individual with locomotion.	Mouse
Allogrooming	Grooming behavior directed towards another individual.	Mouse
Anogenital sniffing	Sniffing and exploring the anogenital (perianal and genital) region of another individual.	Mouse
Back touching	One individual touching the back of another individual by nose or fore limbs.	Mouse
Chasing contact	One individual running after another individual with discrete contact.	Mouse
Immobility	All of the individuals are motionless.	Mouse
Independent close grooming	All of the individuals are self-grooming without the influence of another individual.	Mouse
Independent rearing exploring	All of the individuals are rearing towards the outside without the influence of another individual.	Mouse
Locomotion rearing exploring	All of the individuals are rearing towards the outside with locomotion.	Mouse
Leaving back to back	All of the individuals towards back to each other and in locomotion.	Mouse
Micromovement	Small, subtle movements of individuals.	Mouse
Peer sniffing	Sniffing behavior directed towards another individual.	Mouse

Peer locomotion sniffing	Sniffing behavior directed towards another locomotion individual.	Mouse
Peer grooming sniffing	Sniffing behavior directed towards another grooming individual.	Mouse
Synchronous behavior	The coordinated movement or activity of multiple individuals at the same time.	Mouse
Anogenital touching	Contacting the genital or anal region of another individual.	Bird
Clamping on rectrix	One individual holding onto the tail feathers of another individual by feet.	Bird
Kissing	Two individuals touching beaks.	Bird
Back touching	One individual touching the back of another individual by nose or fore limbs.	Dog
Chasing	One individual pursuing or running after another individual.	Dog
Nose touching	Two individual touching the noses of each other.	Dog

Extended Data Tab. 1| Social behavior definitions for manual labeling. The definition of social behavior of mouse, bird and dog refers to Mouse Ethogram database (www.mousebehavior.org), ref.^{35,51–54}.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [figSM1.ai](#)
- [figSM2.ai](#)
- [figSM3.ai](#)
- [flatPWeispc.pdf](#)
- [flatPWeiepc.pdf](#)
- [flatPWeirs.pdf](#)