Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the supplemental impact statement (Appendix 2).
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see Appendix 5.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, this information is included in the Results and in Appendix 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We did not run experiments multiple times, but we reported multiple measures of the statistical distributions of error metrics in Table 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This information is reported in Appendix 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A] Yes, the CC BY 4.0 license is included in the datasheet (Appendix 1) and in our repositories.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, the dataset is available at figshare as detailed in the Datasheet (Appendix 1).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We collected the data ourselves from animals, following protocols for animal care approved by the Harvard University IACUC (Appendix 2).
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data come from animals and thus have no personally identifiable information of offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix 1 | Dataset Nutrition Label

PAIR-R24M Datasheet

Dataset PAIR-R24M
Number of Frames 24 Million
Number of Annotated Animals 7
Number of Unannotated Animals 2

Metadata

Filename	README.txt
Format	.mp4, .csv, .json
URL	https://figshare.com/articles/dataset/PAIRS_dataset/14754374
DOI	https://doi.org/10.6084/m9.figshare.14754374.v2
Keywords	Animal Behavior, Pose Estimation, Social Behavior
Rows	Timepoints
Columns	3D keypoint positions, behavior labels
Missing Data	Stored as NaN
License	CC BY 4.0
First Released	June 7 2021

Variables | markerDataset

`center_of_mass`	Center of mass of the animal
`aligned_position`	Marker positions aligned to center of mass
`absolute_position`	Marker positions in global arena coordinates
`goodFrame`	Frames without missing markers
`behavior`	Behavior of the animal
`interactionCategory`	Interaction category of the animal pair

Variables | Calibration

`rotationMatrix`	Rotation Matrix of Camera
`translationMatrix`	Translation Matrix of Camera
`intrinsicMatric`	Intrinsic Matrix of Camera
`radialDistortion`	Rotational Distortion Coefficient
`tangentialDistortion`	Translational Distortion Coefficient

Figure 4: PAIR-R24M nutrition label, constructed using the template from Bandy et al. [66]. Note that the exact DOI may change as the dataset is updated.

Appendix 2 | Impact and Animal Care Statement

Preclinical screening of animal models is a crucial step in the drug discovery pipeline, and developing improved social assays thus represents an important step to alleviating human disease burden. Careful measurements and associated analysis frameworks to understand the natural behavior of animals in 3D should facilitate new approaches for animal phenotyping and contribute to the development of new theraputics, especially in the cases of neuropsychiatric diseases that affect social behaviors, such as Autism Spectrum Disorders, Williams syndrome, and schizophrenia. All experiments were performed at Harvard's AAALAC-accredited animal facility. The care and experimental manipulation of all animals were reviewed and approved by the Harvard University Faculty of Arts and Sciences Institutional Animal Care and Use Committee. All surgical procedures were designed to limit pain and discomfort. More details on the experimental procedures are given in [52].

Appendix 3 | Discrepancies in Motion Capture and Video Tracking

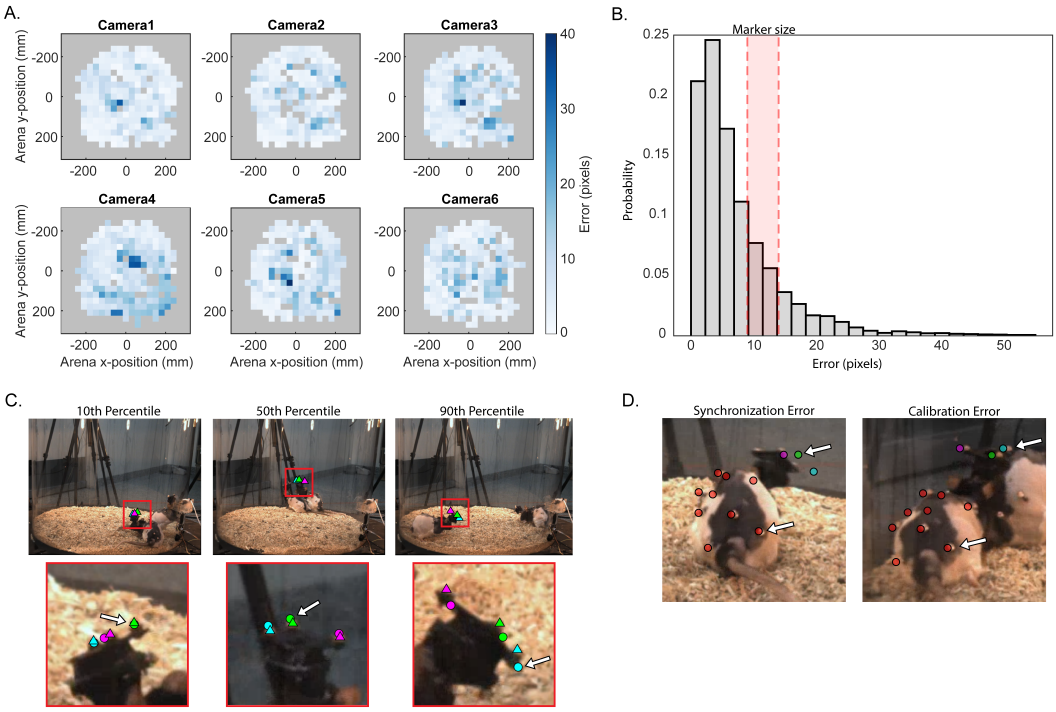

Figure 5: (A) 2D histograms of the discrepancy in the position of the three markers on an animal's head between hand labels and motion capture. Heatmaps depict the differences in pixels between the projections of the motion capture data into 2D and the human hand-labeled points as a function of the x- and y- position of the headcap markers in the arena, for a single camera view. (B) Histogram of the discrepancy in pixels across all cameras in all views. Errors range from 0.04 to 50.1 pixels (px), with a mean error of 7.1 px (or around 2.7 mm). The exact pixel size of the retroreflective marker (5 mm in diameter) depends on the camera view and is indicated by the shaded red bar. (C) Three example frames (top) showing the 10th percentile (left; error = 1.3 px), 50th percentile (center; error = 4.8 px), and 90th percentile (right; error = 15.7 px) discrepancy from camera 5 in (A). The white arrows in the zoomed in images (bottom) highlight the marker representing the respective percentile. (D) An example of synchronization error (left) and calibration error (right) with arrows pointing to a head marker and a body marker for comparison. In the frame with synchronization error, the head markers show larger error than the body markers, likely due to the animal moving its head quickly and the RGB video lagging behind. In the frame with calibration error, the head and body marker errors are more uniform, making an issue with calibration parameters more likely.

In a subset of video frames and camera views, we observed a discrepancy between the marker positions, as tracked using motion capture, and the apparent marker positions in the video frames. Such a discrepancy could be caused by either noise camera calibration, or temporally localized variability in RGB video camera synchronization with motion capture. To quantify the magnitude and extent of these discrepancies, we hand labeled the position of the markers on the head in 2078 video frames and compared them with the projections of points tracked using motion capture. Differences varied across cameras and positions of the animal in the arena (Fig. 5A). On average, differences (7 px mean, 5 px median) were well below both the marker size (9-14 px) and measured precision of hand-labelers (12 px [52]; Fig. 5B-C). Nevertheless, on 10% of frames these differences were greater than the marker diameter, although they rarely exceeded two marker diameters ($\sim 1\%$ of frames). Motion capture discrepancies appeared notably smaller for markers on the body, which are less sensitive to slight variability in synchronization (Fig. 5D). Discrepancies are nearly unavoidable in large datasets [38], and can in principle add robustness to 3D markerless pose detection models [18]. Nevertheless, these deviations may present a noise ceiling for 3D pose tracking, and could be removed, if desired, when running benchmarks [38].

Appendix 4 | Constant Head Segment Lengths Suggest Accurate Animal Identity Tracking

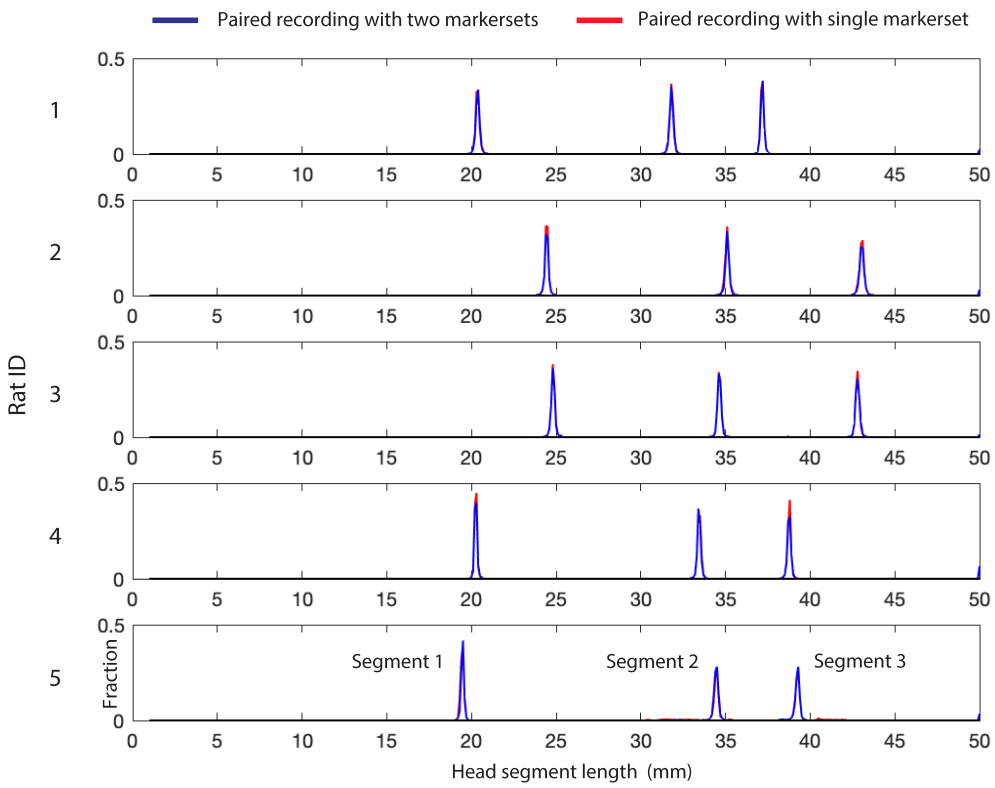

Figure 6: Normalized histograms of head segment lengths for Subjects 1-5, measured from all recorded motion capture data and broken down by recording type: paired recordings in which only one subject had markers (red lines) and paired recordings in which both subjects had markers (blue lines). For each subject, histograms for each of the three head segments are plotted together on one graph.

Motion capture measurements are so precise that they enable fingerprinting of each subject via quantification of small subject-specific differences in head segment lengths; these differences arise from variability in marker placement during headcap construction. We established reference head segment lengths for each subject by examining their distributions in marker + markerless recordings, where identity swapping is impossible. In marker + marker recordings, swaps in animal identity should manifest as frames with head segment lengths deviating from each animal's reference. We see little support for such swaps in the data.

Appendix 5 | DANNCE Training and Evaluation

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11
DANNCE.L2*	7.82	8.24	8.10	8.08	8.84	8.24	8.25	8.23	10.16	8.12	8.21
DANNCE.L2	7.08	6.53	6.96	7.21	7.31	6.67	7.64	7.45	9.82	7.27	7.08
DANNCE.L1*	**6.28**	**6.03**	**6.21**	**6.40**	**6.58**	**5.87**	**6.78**	**6.49**	9.78	6.38	6.34
DANNCE.L1	6.46	6.19	6.26	6.56	6.77	6.12	6.97	6.74	**9.39**	**6.32**	6.34

Table 3: MPJPE in validation subject 5, broken down by individual behavioral category.

	IB1	IB2	IB3
DANNCE.L2*	8.68	9.30	8.57
DANNCE.L2	7.44	8.24	7.26
DANNCE.L1*	**6.61**	**7.22**	6.56
DANNCE.L1	6.74	7.58	**6.47**

Table 4: MPJPE in validation subject 5, broken down by interaction behavioral category.

Multi-animal DANNCE (https://github.com/spoonsso/dannce/) training and evaluation was performed in Python 3.7 using tensorflow (for the network) and pytorch (for parallel 3D volume generation). For efficiency, we trained multi-animal DANNCE using 4 NVIDIA V100 16 GB GPUs on the Harvard Odyssey compute cluster. We used training frames and ground-truth poses from 4 unique animal pairs, distributed over 7 1-hour recordings at 30 Hz. To form the training set, 10,000 time points (60,000 frames) were sampled randomly without replacement from the time points in each recording having a complete motion capture marker set without imputation, resulting in 70,000 training samples total. We chose at the outset to train each DANNCE network for 30 epochs using a batch size of 4, and at the end of training we evaluated the performance of each network on the full validation dataset just once (results in Table 2, 3, 4). For the benchmarks presented here, we used all samples from a 1-hour recording of subject 3 and 5, evaluated over withheld validation subject 5 only, that had a complete motion capture marker set without imputation (43,285 samples; 259,710 frames). For each animal, we anchored its image volume to the 3D position of its "SpineM" marker in each frame.

For the benchmarks, we varied the loss function used for DANNCE training, using either mean squared error (L2) or mean absolute error (L1). We also tested training DANNCE from a random weight initialization, or from previously published weights found by training over images of single animals behaving in the Rat 7M dataset (https://github.com/spoonsso/dannce/) [18]. In all cases, we used DANNCE in the "AVG" architecture configuration (a 3D U-Net with a soft-argmax output layer) and trained using the Adam optimizer with lr = 0.001 and default parameters. We list the full set of DANNCE training parameters used in Table 5. Full architecture details and parameter definitions can be found on the dannce github.

To quantify DANNCE performance, we calculated standard 3D pose estimation error metrics, using a Procruste's alignment to ground-truth before calculations (translation and rotation only; no scaling). MPJPE was calculated as the mean Euclidean error across all markers after alignment. $PJPE_{50}$ is the median error across all markers. PCK metrics reflect accuracy over all markers after binarizing all predictions using the indicated threshold distances, expressed as fractions of the distance between two Head markers (19.4 mm). For the mPCK metric, we calculated PCK for each threshold in [0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1] and took the mean.

Parameter	Value
nvox	64
n_channels_in	3
n_views	6
n_channels_out	20
new_last_kernel_size	[3 3 3]
batch_size	4
epochs	30
loss	'mask_nan_keep_loss','mask_nan_l1_loss'
lr	'1e-3'
net	'unet3d_big_expectedvalue'
n_layers_locked	0
num_train_per_exp	10000
vmin	-120
vmax	120
interp	'nearest'
rotate	1
expval	1
channel_combo	'None'
n_rand_views	6
predict_mode	'torch'
data_split_seed	11516
depth	0
augment_continuous_rotation	0
mono	0
augment_hue	0
drop_landmark	'None'
raw_im_h	1048
raw_im_w	1328
mirror	0
n_instances	1
write_npy	'None'

Table 5: Values of DANNCE training parameters.