SUPPLEMENTARY A

Preliminaries. ST-GCN [37] is the first work adopting Graph Convolutional Networks for skeleton data modelling. It is constructed by stacked spatio-temporal blocks, each of which is composed of a spatial convolution (GCN) block, followed by a temporal convolution (TCN) block. The spatial module utilizes the GCN to model the structural dependencies of nodes, which is formulated as:

$$\mathbf{X}_t^{(l+1)} = \sum_k^{K_v} \mathbf{W}_k \left(\mathbf{X}_t^l \mathbf{A}_k\right) \tag{1}$$

where K_v denotes the kernel size. l is the layer index of the GCN. $\mathbf{W}_k$ is a trainable weight matrix that is implemented as $C_{\text{out}} \times C_{\text{in}} \times 1 \times 1$ convolution operation, where C_{out} and C_{in} are the output and input channels. $\mathbf{A}_k = \mathbf{\Lambda}_k^{-\frac{1}{2}} \tilde{\mathbf{A}}_k \mathbf{\Lambda}_k^{-\frac{1}{2}}$, where $\tilde{\mathbf{A}}_k$ is the adjacency matrix of the skeleton graph indicating intra-skeleton connections. $\mathbf{\Lambda}_k$ is the diagonal matrix, where $\mathbf{\Lambda}_k^{ii} = \sum_j \left(\tilde{\mathbf{A}}_k^{ij}\right) + c$, and c is a small constant avoiding empty rows. On the temporal dimension, TCN is implemented by applying a $K_t \times 1$ 2D convolution operation to the input $\mathbf{X} \in \mathbb{R}^{C \times T \times N}$ with (T, N) dimensions, where K_t is the kernel size.

The structure of the skeleton graph shown in Eq. (1) is predefined by a fixed adjacency matrix. In order to learn an adaptive topology, [26] presented the Adaptive Graph Convolutional Network (A-GCN), in which the adjacency matrix is divided into three complementary parts, as shown in Eq. (2):

$$\mathbf{X}_t^{(l+1)} = \sum_k^{K_v} \mathbf{W}_k \mathbf{X}_t^l \left(\mathbf{A}_k + \mathbf{B}_k + \mathbf{C}_k\right) \tag{2}$$

where $\mathbf{A}_k$ is the same as the one shown in Eq. (1), which represents the physical structure of human body. $\mathbf{B}_k$ can be learned according to the training data and its elements can be an arbitrary value. It indicates the existence and strength of the connections between two nodes. $\mathbf{C}_k$ determines the connection strength between two nodes by calculating their similarity using the normalised embedded Gaussian function.

Algorithm S1 Interaction-aware Transformer (IAT) reasoning process.

Input: Spatial-temporal node-level representation embedded with intra-, inter- and cross-skeleton interactions, i.e., $\mathbf{X}_{s_i}^{\text{cross},l}$ shown in Eq. (12).

Output: Graph-level representation $\widetilde{\mathbf{X}}_g^{\text{cross}}$ for social behaviour classification.

for $l \leftarrow 1$ to L **do**
  for $i \leftarrow 1$ to I **do**
    for $j \leftarrow 1$ to $J-1$ **do**
      $\mathbf{H}_{s_{ij+1},m_{j+1}}^{\text{cross},l} \leftarrow LN(\mathbf{Q}_{s_{ij+1},m_{j+1}}^l + \mathbf{\Upsilon}_{m_{j+1}}(\mathbf{X}_{s_{ij},m_j}^{\text{cross},l}))$ (Eq. (13));
      $\mathbf{X}_{s_{ij+1},m_{j+1}}^{\text{cross},l} \leftarrow LN(\mathbf{H}_{s_{ij+1},m_{j+1}}^{\text{cross},l} + \mathbf{\Gamma}(\mathbf{H}_{s_{ij+1},m_{j+1}}^{\text{cross},l}))$ (Eq. (14));
    end for
    $SAP(\mathbf{X}_{s_{i1}}^{\text{cross},l}) \leftarrow \frac{1}{N_{s_i}^1} \sum_{m_1=1}^{N_{s_i}^1} \mathbf{X}_{s_{i1},m_1}^{\text{cross},l}$;
    $SMP(\mathbf{X}_{s_{i1}}^{\text{cross},l}) \leftarrow max_{m \in \mathcal{N}_{s_i}^1}(\mathbf{X}_{s_{i1},m_1}^{\text{cross},l})$;
    $IAT(\mathbf{X}_{s_{i1}}^{\text{cross},l}) \leftarrow \mathbf{X}_{s_{iJ},m_J}^{\text{cross},l}$ (w.r.t Eqs. (13) and (14));
  end for
  $\widetilde{\mathbf{X}}_g^{\text{cross},l} \leftarrow IAT([SAP(\mathbf{X}_{s_{11}}^{\text{cross},l}); \cdots; IAT(\mathbf{X}_{s_{I1}}^{\text{cross},l})])$ (Eq. (3));
  if $l < L$ **then**
    $\widetilde{\mathbf{X}}_{s_{iJ}}^{\text{cross},l} \leftarrow IAT(\widetilde{\mathbf{X}}_g^{\text{cross},l})$ (4));
    $\mathbf{X}_{s_i}^l = \widetilde{\mathbf{X}}_{s_{iJ}}^{\text{cross},l} + \mathbf{X}_{s_i}^{\text{cross},l}$;
    $\mathbf{X}_{s_i}^{\text{cross},(l+1)} = CS - NLI(\mathbf{X}_{s_i}^l)$ (w.r.t Eqs. (1), (10) and (12));
  end if
end for
$\widetilde{\mathbf{X}}_g^{\text{cross}} \leftarrow [\widetilde{\mathbf{X}}_g^{\text{cross},1}; \widetilde{\mathbf{X}}_g^{\text{cross},2}; \cdots; \widetilde{\mathbf{X}}_g^{\text{cross},L}]$
return Final graph-level representation $\widetilde{\mathbf{X}}_g^{\text{cross}}$.

More details about graph-level representation enhancement in IAT. Given the representation of the first subgraph on the l-th layer of our network, i.e., $\mathbf{X}_{s_{i1}}^{\text{cross},l}$, we first calculate the average and maximum values of the representation in the spatial domain by spatial average pooling [26] $SAP(\cdot)$ (see Algorithm S1) and max pooling [34] $SMP(\cdot)$, respectively. Eqs. (13) and (14) can be treated as the implementation of function $IAT(\cdot)$ that describes the graph-level representation. Instead of fusing different graph-level representations across skeletons by direct summing over the spatial dimension, we attempt to model

the relations between them using our proposed interaction-aware self-attention module to adaptively enhance the graph-level representation, formulated as follows:

$$\begin{aligned}\widetilde{\mathbf{X}}_{g}^{\text{cross},l} =& IAT(\mathbf{Z}_{s_{i1}}^{\text{cross},l}) \in \mathbb{R}^{C_l \cdot T_l} \\ \mathbf{Z}_{s_{i1}}^{\text{cross},l} =& [SAP(\mathbf{X}_{s_{11}}^{\text{cross},l}); SMP(\mathbf{X}_{s_{11}}^{\text{cross},l}); \\ & \cdots ; IAT(\mathbf{X}_{s_{21}}^{\text{cross},l})] \in \mathbb{R}^{6 \times C_l \cdot T_l}\end{aligned} \tag{3}$$

where $\widetilde{\mathbf{X}}_{g}^{\text{cross},l}$ is the enhanced graph-level representation that fuses various semantic information. $\mathbf{Z}_{s_{i1}}^{\text{cross},l}$ is the fused representation, including 3 types of graph-level representation at each skeleton branch.

More details about decoder in IAT. In most existing work [26], [37], the node-level representation of one GCN-TCN block is directly fed into the next block for deeper spatio-temporal representation encoding, where the graph-level representation can be generated based on the last node-level representation. Different from these standard work, we add a decoder to the end of the encoder to adaptively update the node-level representation before sending the representation to the next layer. We directly infer the node-level representation from the graph-level representation using our proposed interaction-aware self-attention presented in Section III-B1:

$$\widetilde{\mathbf{X}}_{s_{iJ}}^{\text{cross},l} = IAT(IAT(\mathbf{Z}_{s_{i1}}^{\text{cross},l})) \in \mathbb{R}^{N_{s_i} \times C_l \cdot T_l} \tag{4}$$

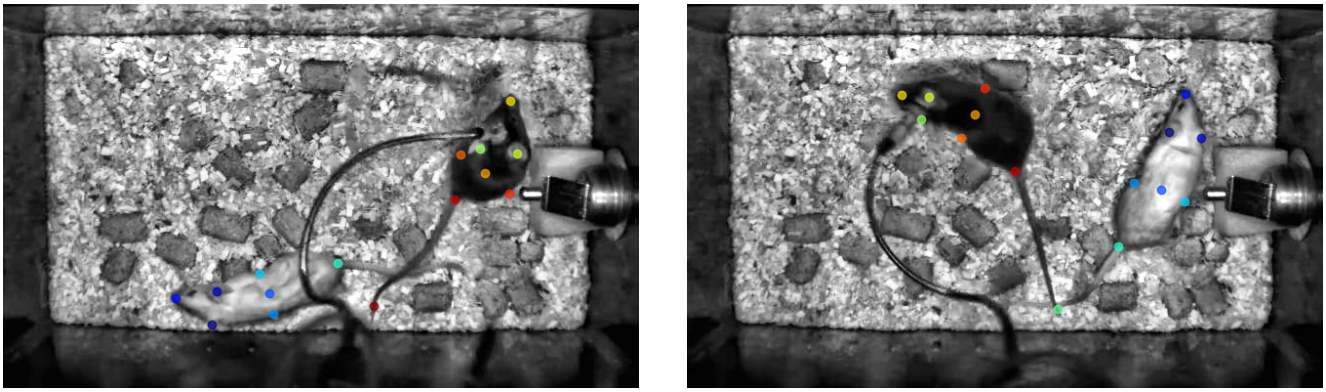
where we define J subgraphs for the decoder and the last one is $\widetilde{\mathbf{X}}_{s_{iJ}}^{\text{cross},l}$. Hence, the node-level representation for the l-th layer can be updated by $\mathbf{X}_{s_i}^{l} = \widetilde{\mathbf{X}}_{s_{iJ}}^{\text{cross},l} + \mathbf{X}_{s_i}^{\text{cross},l}$.

Classification loss. The classification loss is defined as:

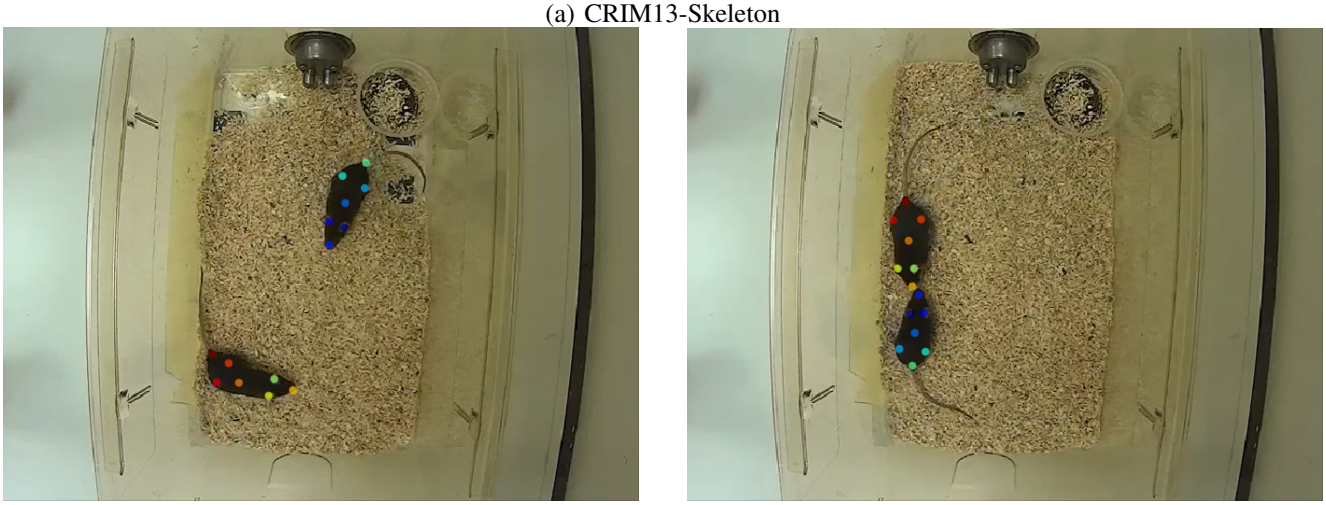
$$\begin{aligned}\widetilde{\mathbf{Y}} &= Softmax(f_o(\widetilde{\mathbf{X}}_{g}^{\text{cross}})) \\ \mathcal{L}_{\text{class}} &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C}\mathbf{Y}_{j}^{(i)} \log \widetilde{\mathbf{Y}}_{j}^{(i)}\end{aligned} \tag{5}$$

where $f_o(\cdot)$ is a fully connected layer. $\widetilde{\mathbf{X}}_{g}^{\text{cross}}$ represents the final representation for classification, which is constructed by concatenating the graph-level representations of different layers. $\widetilde{\mathbf{Y}}_{j}^{(i)}$ represents the predicted probability that the spatio-temporal skeleton graph with feature $\mathbf{X}^{(i)}$ belongs to class j, and $\mathbf{Y}_{j}^{(i)}$ is the corresponding ground truth. B and C denote the numbers of sliding windows and classes, respectively.

SUPPLEMENTARY B



(a) CRIM13-Skeleton



(b) PDMB-Skeleton

Fig. S1. Annotated locations of different mouse keypoints on the CRIM13-Skeleton and PDMB-Skeleton datasets. (a) The public CRIM13-Skeleton dataset [17] contains 8 keypoints for each mouse (i.e., 0-left ear, 1-right ear, 2-snout, 3-centroid, 4-left lateral, 5-right lateral, 6-tail base and 7-tail end). The numbers represent the order which the body-parts were annotated. (b) To establish our PDMB-Skeleton dataset, we extract frames every 500ms for each video, and all the extracted video frames were annotated using a freeware DeepLabCut (available at https://github.com/DeepLabCut/DeepLabCut). A team of five professionals were trained to annotate the keypoints of each mouse. Similar to the CRIM13-Skeleton dataset, we annotate the locations of 7 body parts (i.e., 0-left ear, 1-right ear, 2-snout, 3-centroid, 4-left hip, 5-right hip and 6-tail base) for each mouse. We do not annotate the tail end because this keypoint is often occluded and the mouse tail is highly deformable in videos. We only use 7 body parts in all experiments on both PDMB-Skeleton and CRIM13-Skeleton datasets. In particularly, we ensure that the identity of each mouse remains unchanged during the process of annotation.

Dataset Construction. We adopted a careful dataset construction process for our PDMB dataset (see Fig. S1). Specifically, we did not directly utilise a pre-trained model from DeepLabCut [13] to generate keypoint positions. Instead, we took the following steps: (1) Partial Frame Extraction: We initially extracted partial frames (at intervals of 500ms) from each video in the dataset. (2) DeepLabCut Annotation: The selected frames were then annotated using the DeepLabCut tool to manually label the positions of keypoints. (3) PDMB Training Set Construction: Subsequently, we utilised these annotated frames to construct the training set for PDMB. (4) Pose Estimation Network Training: DeepLabCut was trained for mouse pose estimation on the PDMB dataset. (5) Keypoint Estimation: Finally, we used the pretrained pose estimation model to generate keypoint data (including confidence scores) for every frame in the dataset.

Despite the presence of estimation errors, each keypoint is associated with a confidence score that quantifies this error. This confidence score is leveraged as a feature for each keypoint during network training, allowing the model to account for and learn from the uncertainties in the keypoint positions. This approach aligns with methodologies similar to SimBA [17], which also utilises DeepLabCut for keypoint labelling. Additionally, in the case of the public dataset CRIM13-Skeleton, a confidence score is also included to measure position errors for each keypoint.

Data Annotation. Unlike the annotation method for CRIM13-Skeleton, we chose to annotate the left hip and right hip positions for the PDMB-Skeleton dataset. We referenced MARS [36] for mouse keypoint location annotation, labelling the left hip and right hip positions. The reason behind this choice is that the hip positions serve as crucial connectors between the upper body and the tail of the mouse. Considering the holistic perspective, the distribution of these seven keypoints, including

the hip positions, is expected to provide a more comprehensive representation of the mouse's overall body structure. This can potentially contribute to a more nuanced understanding of mouse behaviour.

It's worth noting that the lack of a standardized mouse keypoint annotation scheme in the research community, including questions about which positions to annotate and the number of keypoints to include, poses a challenge. Different keypoint configurations may impact behaviour analysis. We acknowledge this limitation and plan to deal with this problem in our future work.

SUPPLEMENTARY C

TABLE S1
ETHOGRAM OF THE OBSERVED BEHAVIOURS, DERIVED FROM CRIM13 [63]

Behaviour	Description
approach	Moving toward another mouse in a straight line without obvious exploration.
attack	Biting/pulling fur of another mouse.
copulation	Copulation of male and female mice.
chase	A following mouse attempts to maintain a close distance to another mouse while the latter is moving.
circle	Circling around own axis or chasing tail
drink	Licking at the spout of the water bottle
eat	Gnawing/eating food pellets held by the fore-paws.
clean	Washing the muzzle with fore-paws (including licking fore-paws) or grooming the fur or hind-paws by means of licking or chewing.
human	Human intervenes with mice.
sniff	Sniff any body part of another mouse.
up	Exploring while standing in an upright posture.
walk away	Moving away from another mouse in a straight line without obvious exploration.
other	Behaviour other than defined in this ethogram, or when it is not visible what behaviour the mouse displays.

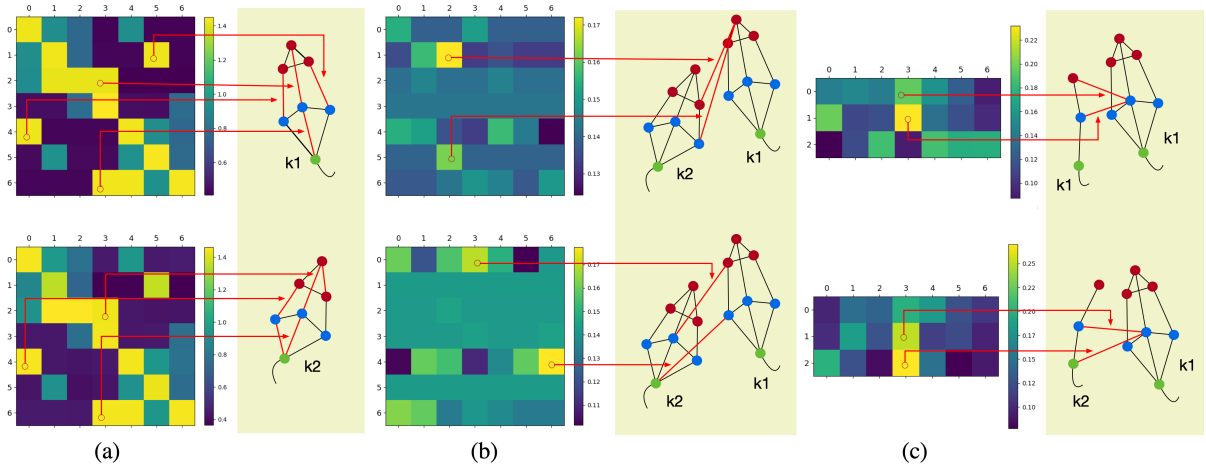
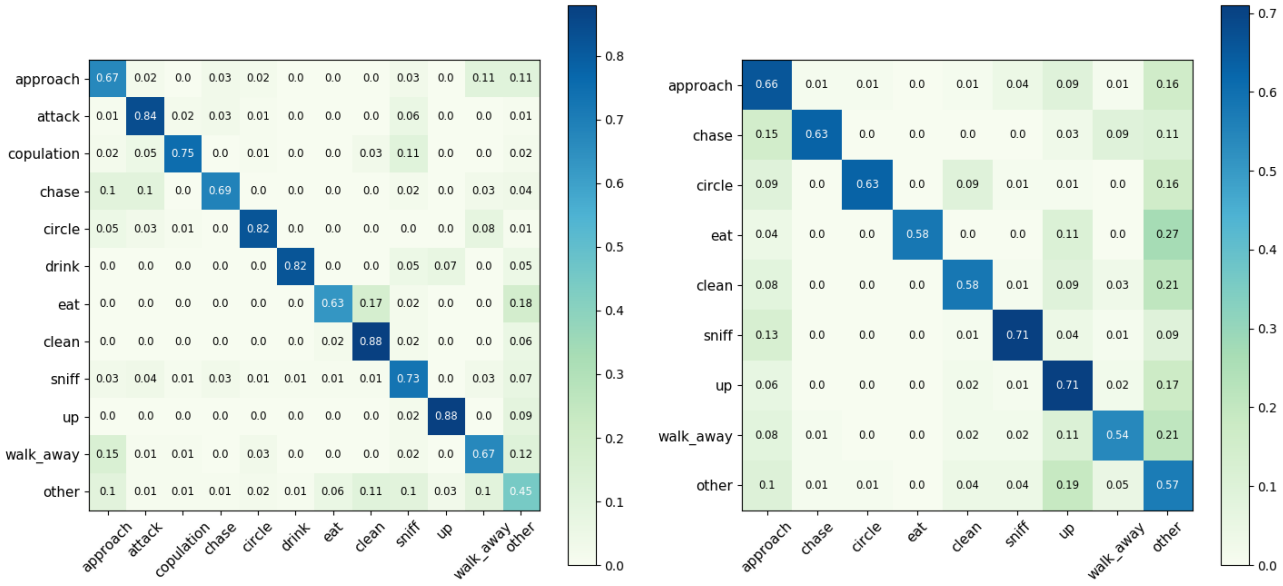


Fig. S2. Visualisation of the learned topologies of a social behaviour sample 'approach' on the CRIM13-Skeleton dataset (at the beginning of training, i.e., epoch=10). (a) The topologies representing intra-skeleton interactions of mouse k_1 (top) and k_2 (bottom). The number of keypoints V is 7 and its configuration is shown in Fig. S1. Here, we show the summation of the learned topologies on the three subsets generated by the partition strategy [37]. (b) The topologies of bidirectional inter-skeleton interactions learned by our model, i.e., $\mathbf{A}^{l=1}_{k_1 \to k_2}$ (top) and $\mathbf{A}^{l=1}_{k_2 \to k_1}$ (bottom). (c) The topologies of cross-skeleton interactions (s_1 to s_2) learned by our model, i.e., $\mathbf{A}^{l=1}_{s_1 \to s_2}$ (top) and $\widetilde{\mathbf{A}}^{l=1}_{s_1 \to s_2}$ (bottom). For each type, we use red lines to indicate the interactions with high significance. We observe that the module generates relatively dense fully connected graph at the beginning of training, especially for the inter- and -cross interactions, i.e., interactions not related to behaviours. On the contrary, our final module (Fig. 5) tends to give less attention to trivial interactions.



(a) CRIM13-Skeleton (b) PDMB-Skeleton

Fig. S3. Confusion matrices of our method (i.e., CS-IGANet) on CRIM13-Skeleton (a) and PDMB-Skeleton datasets (b). The diagonal cells in each confusion matrix show the percentage of correct classifications. The confusion matrix is obtained for measuring the agreement between the ground-truth (row) and the predicted labels (column). The non-diagonal cells contain the percentages of the incorrectly classified behaviors. In each row, all the values should sum to be 1. The higher probabilities along the diagonal and the lower off-diagonal values indicate the degrees of successful classification for all the categories. The colour bar indicates the degree of the agreement whilst deep blue indicates the agreements close to 100%.

TABLE S2
ABLATION EXPERIMENTS FOR THE CROSS-SKELETON NODE-LEVEL INTERACTION (CS-NLI) MODULE ON THE CRIM13-SKELETON DATASET. WE PRESENT THE CLASSIFICATION ACCURACY (%) OF EACH BEHAVIOUR, AVERAGE ACCURACY OVER ALL THE BEHAVIOURS, FLOPS AND PARAMETER NUMBER. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD.

Methods	Approach	Attack	Copulation	Chase	Circle	Drink	Eat	Clean	Sniff	Up	Walk away	Other	Average	Params	FLOPs
CS-NLI(two dense graphs)	63.35	77.91	72.02	40.91	59.79	**79.93**	**62.35**	83.52	62.73	84.65	55.75	**50.67**	66.13	3.41M	0.40G
CS-NLI(multi-scale graphs)	**69.24**	**81.81**	**78.91**	**44.73**	**66.23**	79.12	57.70	**86.87**	**65.72**	**85.12**	**59.10**	45.02	**68.30**	2.90M	0.37G

TABLE S3
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE NTU-INTERACTION AND NTU120-INTERACTION DATASETS IN ACCURACY (%).

Method	NTU-Interaction		NTU120-Interaction	
	X-Sub	X-View	X-Sub	X-View
ST-GCN* [37]	89.31	93.72	80.69	80.27
2S-AGCN* [26]	93.36	96.67	87.83	89.21
CTR-GCN* [39]	95.31	97.60	92.03	92.82
2S-DRAGCN* [27]	94.68	97.19	90.56	90.43
2P-GCN* [28]	97.05	**98.80**	93.47	93.73
Ours+ (CS-IGANet)	**97.12**	97.89	**94.39**	**94.70**

* Results are reported in [28].
+ We follow the same pre-processing method as described in [28]. We adopt multi-scale skeleton graphs composed of 25 joints (for each actor) and 12 joints [32], respectively, to serve as dense and sparse skeleton graphs in our framework.

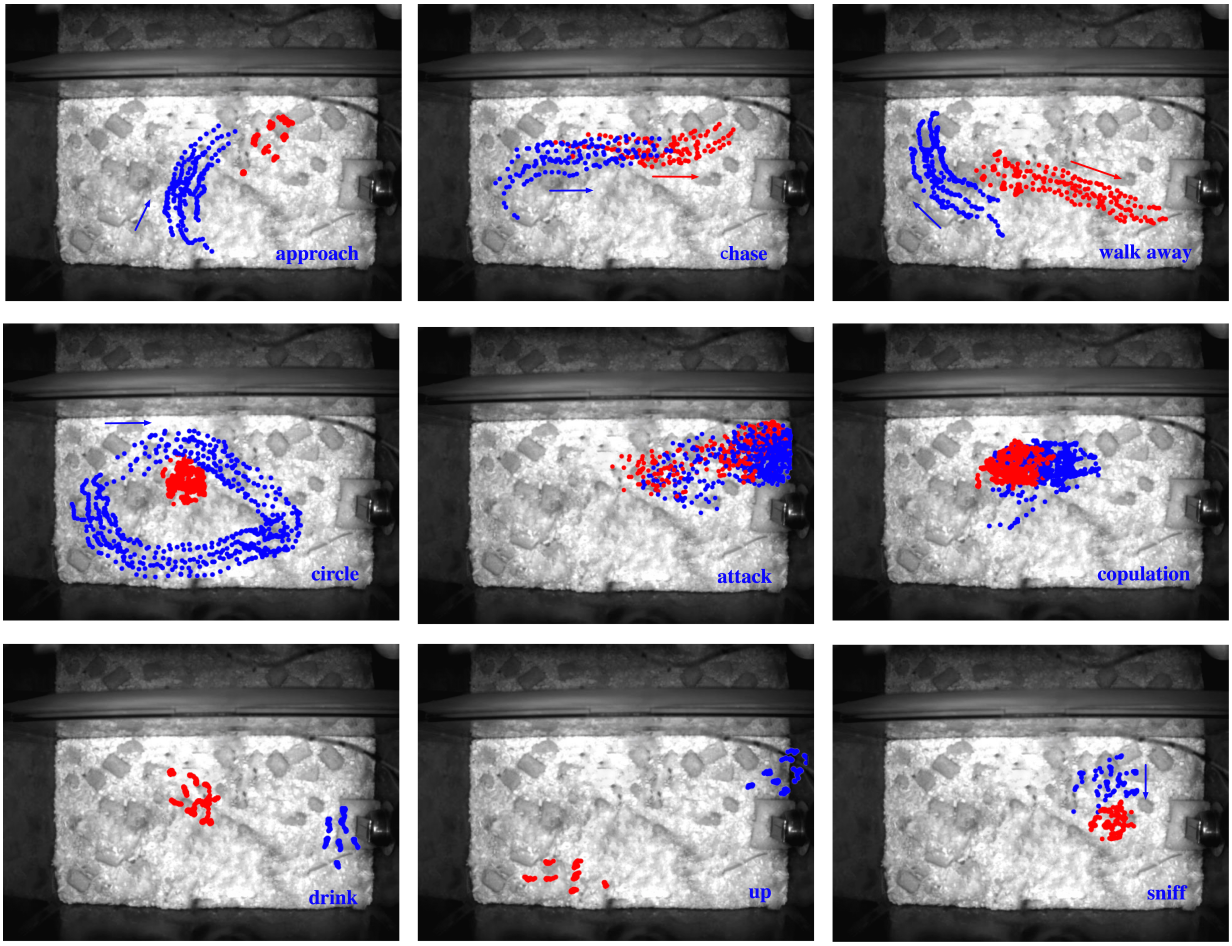


Fig. S4. Examples of motion trajectory of different behaviours in the CRIM13-Skeleton dataset. Blue and red points indicate the keypoints of the resident mouse and the intruder, respectively. Blue and red arrows represent the direction of motion. For some behaviours (e.g., clean and up) without significant movement, we do not give the direction of motion.