SUPPLEMENTARY A

Additional qualitative results. As shown in Fig. S1, we visualise the I3D feature and other features learned by ICC and our method using t-Distributed Stochastic Neighbor Embedding (t-SNE) on GTEA datasets. Different colours represent different actions. Our approach leads to better separation of different classes, demonstrating the strong representation learning ability of our Semantic-guided Multi-level Contrast scheme.

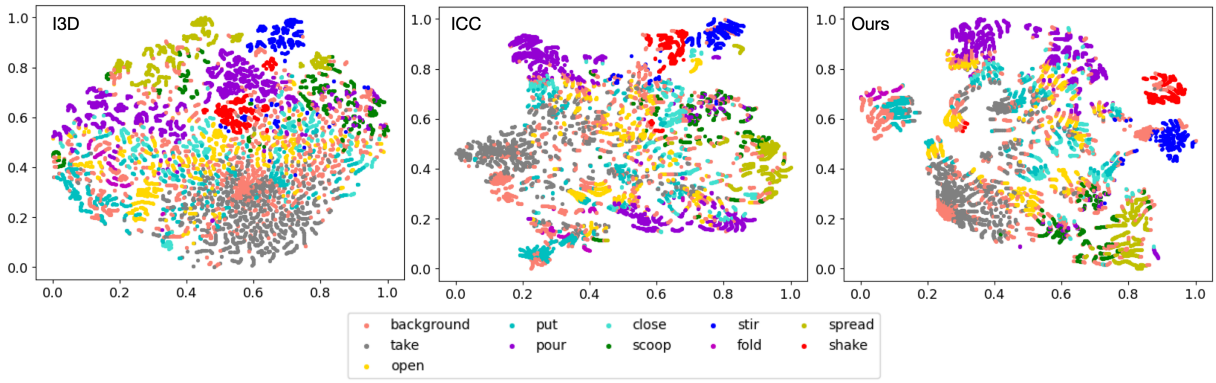


Fig. S1. t-SNE visualisation of the I3D feature and other features learned by ICC and our method. Each point represents an image frame. We show all behaviour classes (11) of the GTEA dataset in different colours.

Compare SMC with other unsupervised representation learning methods. In Tab. 3 (main paper) of the paper, we have compared our unsupervised method with that of ICC. Since ICC is the first work based on unsupervised representation learning for semi-supervised action segmentation, so far we could not find any other publications exploring unsupervised representation learning for this task. Thus, we attempt to compare our unsupervised method against a SOTA method [1] of unsupervised video representation learning, but it is not quite related to this task. We use the same positive and anchor representations as those of our work but the negative representation is obtained by temporal shuffling [1] of anchor representation. From Tab. S1, our method shows better performance.

TABLE S1
COMPARING OUR SMC WITH SCVRL [1].

Dataset	Method	F1@{10, 25, 50}			Edit	Acc
50Salads	SCVRL [1]	48.0	43.5	34.4	37.8	69.2
	SMC	**58.1**	**54.0**	**43.5**	**46.3**	**75.1**
GTEA	SCVRL [1]	72.5	66.0	48.6	64.9	71.2
	SMC	**78.9**	**74.3**	**59.2**	**73.0**	**76.2**
Breakfast	SCVRL [1]	46.8	41.9	31.8	38.6	71.1
	SMC	**59.7**	**55.4**	**42.8**	**52.7**	**72.1**

TABLE S2
TRAINING HYPERPARAMETERS LEARNING RATE (LR), WEIGHT-DECAY (WD), EPOCHS (EPS.) AND BATCH SIZE (BS) OVER DIFFERENT DATASETS FOR UNSUPERVISED AND SEMI-SUPERVISED LEARNING.

Model	Breakfast				50Salads				GTEA				PDMB			
	LR	WD	Eps.	BS	LR	WD	Eps.	BS	LR	WD	Eps.	Bs	LR	WD	Eps.	Bs
(**T**: **S**)	le-3	3e-3	100	50	1e-3	1e-3	100	5	1e-3	3e-4	100	4	1e-3	3e-4	100	4
C	le-2	3e-3	200	50	1e-2	1e-3	400	5	1e-2	3e-4	400	4	1e-2	3e-4	400	4
(**T**:**G**:**S**)	le-5	3e-3	200	50	1e-5	1e-3	400	5	1e-5	3e-4	400	4	1e-5	3e-4	400	4

SUPPLEMENTARY B

TABLE S3
ETHOGRAM OF THE OBSERVED BEHAVIOURS [2].

Behaviour	Description
approach	Moving toward another mouse in a straight line without obvious exploration.
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.
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.
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	ehaviour other than defined in this ethogram, or when it is not visible what behaviour the mouse displays.

Additional related works about temporal modelling of mouse behaviour. In recent years, temporal dependencies among actions have also been investigated to facilitate mouse behaviour modelling. Jiang et al. [3] employed a Hidden Markov Model (HMM) to model the contextual relationship among adjacent mouse behaviours over time. Specifically, they represented each action clip as a set of feature vectors using spatial-temporal Segment Fisher Vectors (SFV), which were then treated as observed variables in the HMM. In addition, Jiang et al. [2] proposed a deep graphic model to explore the temporal correlations of mouse social behaviours, which demonstrated the advantage of modelling behavioural correlations. However, these methods mainly focus on the correlations between the neighbouring behaviours, which is difficult to capture multi-scale temporal dependencies of mouse behaviours in long videos. Also, these methods usually require fully supervised data, which is obtained by manually annotating the exact temporal location of each behaviour occurring in all training videos. Such data collection is expensive, particularly in behavioural neuroscience [4], where datasets are usually complex and lab-specific.

TABLE S4
COMPONENT-WISE ANALYSIS OF THE UNSUPERVISED REPRESENTATION LEARNING FRAMEWORK WITH A LINEAR CLASSIFIER ON THE PDMB DATASET.

Method	PDMB				
	F1@{10, 25, 50}			Edit	Acc
$\mathcal{L}_{ap}^{P}(\mathbf{M}_{in})$+$\mathcal{L}_{aa}^{N}$	39.0	33.9	21.6	36.5	37.7
$\mathcal{L}_{ap}^{P}(\mathbf{I})$+$\mathcal{L}_{aa}^{N}$	**56.3**	**53.6**	**40.8**	**51.8**	**53.4**
$\mathcal{L}_{ap}^{P}(\mathbf{M}_{in})$+$\mathcal{L}_{ap}^{N}$	38.0	34.2	24.3	35.3	42.2
$\mathcal{L}_{ap}^{P}(\mathbf{I})$+$\mathcal{L}_{ap}^{N}$	**55.9**	**53.5**	**41.0**	**40.1**	**54.3**
Constructing positive pairs by $\boldsymbol{M}_{in}$ or I					
$\mathcal{L}_{ap}^{P}$ + $\mathcal{L}_{aa}^{N}$	56.3	53.6	40.8	51.8	53.4
$\mathcal{L}_{ap}^{P}$ + $\mathcal{L}_{ap}^{N}$	55.9	53.5	41.0	40.1	54.3
$\mathcal{L}_{ap}^{P}$ + $\mathcal{L}_{aa}^{N}$ + $\mathcal{L}_{ap}^{N}$	58.5	56.6	43.6	53.4	58.5
$\mathcal{L}_{ap}^{P}$ + $\mathcal{L}_{aa}^{N}$ + $\mathcal{L}_{ap}^{N}$ + $\mathcal{L}_{pp}^{N}$	**59.6**	**58.0**	**45.9**	**53.8**	**61.3**
Comparing different negative pairs					
w/o dynamic clustering	59.6	58.0	45.9	53.8	61.3
w/ dynamic clustering	**61.1**	**59.6**	**47.2**	**55.5**	**62.5**
Dynamic clustering facilitates contrastive learning					

Mouse Social Behaviour Dataset. Our Parkinson's Disease Mouse Behaviour (PDMB) dataset was collected in collaboration with the biologists of Queen's University Belfast of United Kingdom, for a study on motion recordings of mice with Parkinson's disease (PD) [2]. The neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) is used as a model of PD, which has become an invaluable aid to produce experimental parkinsonism since its discovery in 1983 [5]. All experimental procedures were performed in accordance with the Guidance on the Operation of the Animals (Scientific Procedures) Act, 1986 (UK) and approved by the Queen's University Belfast Animal Welfare and Ethical Review Body. We recorded videos for 3 groups of MPTP treated mice and 3 groups of control mice by using three synchronised Sony Action cameras (HDR-AS15) (one top-view and two side-view) with frame rate of 30 fps and 640*480 resolution. Each group consists of 6 annotated videos and

TABLE S5
PERFORMANCE OF THE NCA MODULE ON OUR PDMB DATASET (10%).

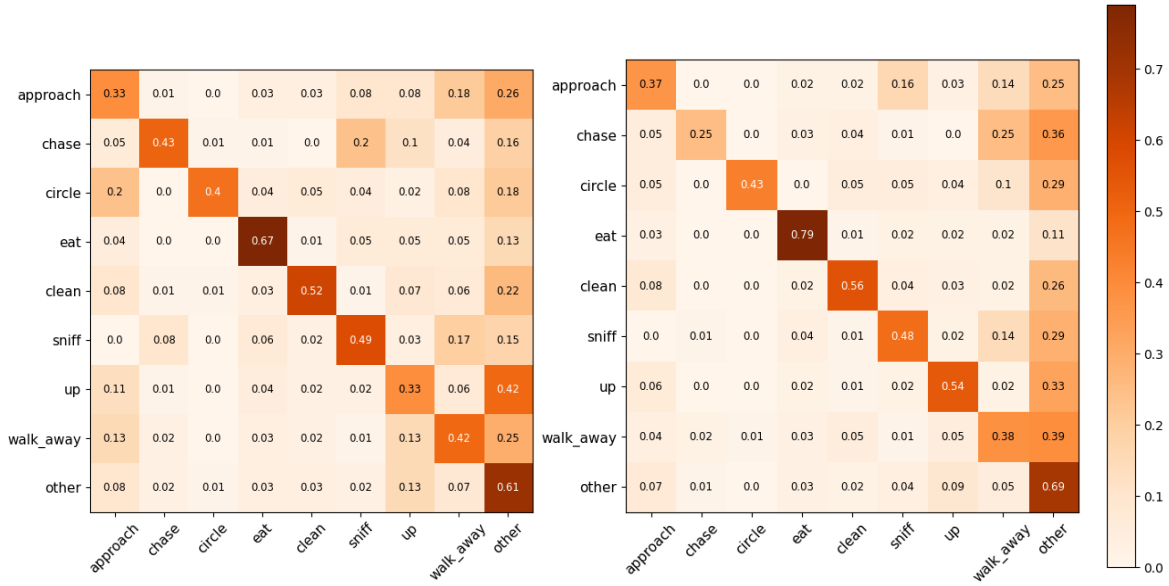
Dataset	Method	F1@{10, 25, 50}			Edit	Acc
PDMB	w/o $\mathcal{L}_{nca}$	54.7	48.7	31.9	45.0	52.7
	w/ $\mathcal{L}_{nca}$	62.0	57.1	40.7	52.7	55.1
	Gain	**7.3**	**8.4**	**8.8**	**7.7**	**2.4**

all videos contain 9 behaviours (defined in Tab. S3) of two freely behaving mice. Different from the experiments of human action segmentation, the input features we use in experiments on this dataset are extracted from the pre-trained model from [6], which encodes the social interactions of mice based on the pose information [7]. The whole dataset is evenly divided into training and testing datasets, and we select 10% or 50% of the videos from the training split for the labelled dataset $\mathcal{D}_L$.

Evaluation of Representation Learning on the PDMB Dataset. As shown in Tab. S4, on the PDMB dataset, utilising $\mathbf{M}_{in}$ would also lead to the generation of pseudo positive pairs, which would impede the efficacy of contrastive learning and consequently result in a significant performance drop. Besides, we achieve the best performance for all metrics when combining three types of negative pairs at the same time, where such combination brings gains of 7.9% and 7% in accuracy for the settings with only $\mathcal{L}_{aa}^{N}$ and $\mathcal{L}_{ap}^{N}$, respectively.

Effect of the NCA Unit on the PDMB Dataset. As shown in Tab. S5, with respect to behavioural correlation modelling of mice, we achieve a significant improvement of more than 7% in F1 and Edit scores on the PDMB dataset.

Finally, to demonstrate the applicability of the proposed framework to behaviour phenotyping of the mice with Parkinson's disease, we investigate the behavioural correlations of both MPTP treated mice and their control strains, as shown in Fig. S2. The findings reveal that MPTP treated mice are more likely to perform 'approach' after 'circle', 'up' or 'walk_away' compared to the control group ('other' is excluded). Besides, MPTP treated mice tend to exhibit 'sniff' behaviour while the normal mice show a higher propensity towards 'walk_away' after 'chase'.



(a) Parkinson's Disease mice (b) Normal mice

Fig. S2. Occurrence frequency of neighbouring behaviours (20-frame interval) for Parkinson's Disease and normal mice. Each cell contains the percentage of the occurrence of behaviour x (along rows) after behaviour y (along column) appears.

REFERENCES

[1] M. Dorkenwald, F. Xiao, B. Brattoli, J. Tighe, and D. Modolo, "Scvrl: Shuffled contrastive video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4132–4141. 1

[2] Z. Jiang, F. Zhou, A. Zhao, X. Li, L. Li, D. Tao, X. Li, and H. Zhou, "Multi-view mouse social behaviour recognition with deep graphic model," *IEEE Transactions on Image Processing*, vol. 30, pp. 5490–5504, 2021. 2

[3] Z. Jiang, D. Crookes, B. D. Green, Y. Zhao, H. Ma, L. Li, S. Zhang, D. Tao, and H. Zhou, "Context-aware mouse behavior recognition using hidden markov models," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1133–1148, 2018. 2

[4] T. D. Pereira, J. W. Shaevitz, and M. Murthy, "Quantifying behavior to understand the brain," *Nature neuroscience*, vol. 23, no. 12, pp. 1537–1549, 2020. 2

[5] V. Jackson-Lewis and S. Przedborski, "Protocol for the mptp mouse model of parkinson's disease," *Nature protocols*, vol. 2, no. 1, p. 141, 2007. 2

[6] F. Zhou, X. Yang, F. Chen, L. Chen, Z. Jiang, H. Zhu, R. Heckel, H. Wang, M. Fei, and H. Zhou, "Cross-skeleton interaction graph aggregation network for representation learning of mouse social behaviour," *arXiv preprint arXiv:2208.03819*, 2022. 3

[7] F. Zhou, Z. Jiang, Z. Liu, F. Chen, L. Chen, L. Tong, Z. Yang, H. Wang, M. Fei, L. Li *et al.*, "Structured context enhancement network for mouse pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2787–2801, 2021. 3