vmTracking: Virtual Markers Overcome Occlusion and Crowding in Multi-Animal Pose Tracking

Hirotsugu Azechi,[1]* Susumu Takahashi[1]

Hirotsugu Azechi: hazechi@mail.doshisha.ac.jp

Susumu Takahashi: stakahas@mail.doshisha.ac.jp

[1]Laboratory of Cognitive and Behavioral Neuroscience, Graduate School of Brain Science, Doshisha University, Kyotanabe, 610-0394, Japan.

***Correspondence: hazechi@mail.doshisha.ac.jp**

Abstract

In multi-animal tracking, addressing occlusion and crowding is crucial for accurate behavioral analysis. Consequently, we introduced Virtual Marker Tracking (vmTracking), which uses virtual markers for individual identification. Virtual markers, created from traditional markerless multi-animal pose tracking tools like multi-animal DeepLabCut (maDLC) and Social LEAP Estimate Animal Poses (SLEAP), attribute features to individuals, enabling consistent identification throughout the entire video without physical markers. Using these markers as cues, annotations were applied to multi-animal videos, and tracking was conducted with single-animal DeepLabCut (saDLC) and SLEAP's single-animal method. vmTracking minimized manual corrections and annotation frames needed for training, efficiently tackling occlusion and crowding. Experiments tracking multiple mice, fish, and human dancers confirmed vmTracking's variability and applicability. These findings could enhance the precision and reliability of tracking methods used in the analysis of complex naturalistic and social behaviors in animals, providing a simpler yet more effective solution.

Keywords: multi-animal tracking, virtual marker, occlusion, crowding, DeepLabCut, SLEAP

Introduction

The study of animal behavior is crucial to advance our understanding of biological, physiological, and neurological processes. Precise simultaneous pose tracking of multiple animals is essential for analyzing naturalistic and social behaviors across species. Recent advancements in machine learning have produced high-precision and versatile single-animal pose tracking tools, such as DeepLabCut (DLC) (1) and LEAP Estimates Animal Poses (LEAP) (2), allowing detailed analysis of complex postures and movements across a wide range of species (3-6). Building on this progress, multi-animal tracking tools such as multi-animal DeepLabCut (maDLC) (7) and Social LEAP (SLEAP) (8) have been introduced. These tools use methods like Kalman filters (9) or optical flow (10) to predict movements and facilitate frame-to-frame identification. However, despite their effectiveness in simpler scenarios, these tools face substantial challenges in multi-animal contexts, particularly with issues such as occlusion, where animals obscure each other, and crowding, where individuals are densely clustered. Despite ongoing research and development efforts to resolve issues of occlusion and crowding (11,12) these problems remain unsolved, hindering accurate multi-animal pose tracking. Currently, research using these tools has succeeded in tracking simple postures like head direction (13-15), but has not yet advanced to accurately tracking detailed postures over extended periods.

When automated systems struggle to achieve high accuracy in animal tracking, human involvement in reviewing, correcting, and providing feedback on the results can enhance accuracy. Thus, human-in-the-loop machine learning, which integrates human judgment into the machine-learning process, can enhance animal-tracking methods (16). Currently, in simple scenarios without occlusion or crowding, accurate multi-animal pose tracking can be achieved with minimal human intervention. However, in situations where occlusion and crowding generate complex interactions, the need for human intervention increases significantly, yet achieving accurate pose tracking remains challenging.

In this study, to address the challenges of occlusion and crowding, we incorporated human intervention during the tracking process to introduce virtual markers into videos of markerless subjects. These markers provide individual identification features without relying on physical markers, which can potentially have unforeseen effects on animal behavior and may inhibit natural behaviors. Through tracking experiments with mice, fish, and humans, we demonstrate that our vmTracking approach with virtual markers enhances the accuracy of machine learning predictions in occlusion and crowding while requiring minimal manual human intervention. The vmTracking provides a robust framework for observing and analyzing social interactions among animals in their natural environments, potentially expanding the boundaries of behavioral science and offering deep insights into collective animal behavior.

Results

vmTracking involves attaching virtual markers to markerless animals to identify and track individuals (Fig. 1). Initially, markerless multi-animal pose tracking was conducted, and the results were visualized

as markers in a video. These were termed 'virtual markers' to distinguish them from physical markers. Videos of multiple animals marked with these virtual markers were then analyzed using a single-animal pose tracking tool (either saDLC or SLEAP's single-animal method, LEAP). During annotation, each animal was labeled with a consistent ID across frames based on the virtual marker. The accuracy of vmTracking was evaluated using either saDLC (vmT-DLC) or LEAP (vmT-LEAP), and compared to traditional markerless multi-animal pose tracking tools (maDLC or SLEAP).

Tracking black mice against black backgrounds

To evaluate the effectiveness of vmTracking in low-contrast environments, where individual identification is particularly challenging due to occlusion and crowding, an experiment was conducted with five C57BL/6J mice on a black background. Each mouse was tracked at six keypoints. Markerless multi-animal pose tracking was performed using maDLC or SLEAP (Fig. 2, A and B). Subsequently, videos were created to visualize two keypoints as virtual markers using SLEAP (Fig. 2, A and C), and vmT-DLC and vmT-LEAP were conducted (Fig. 2D and Supplementary Video 1 and 2). The results stemmed from five iterations of retraining for both maDLC and vmT-DLC, with the final total number of annotated frames being 1,097 and 1,196, respectively. The results for SLEAP and vmT-LEAP were based on training and predictions using annotation data imported from maDLC and vmT-DLC. Our analysis of the four tracking methods revealed significant improvements in all five evaluation metrics (maDLC vs. vmT-DLC: Matches: $p < 0.001$, false negatives (FN): $p < 0.001$, false positives (FP): $p < 0.05$, ID switches: $p < 0.001$, root mean square error (RMSE): $p < 0.05$, Greenhouse–Geisser-corrected, repeated-measures analysis of variance (ANOVA), $n = 12$, Fig. 2E). Specifically, vmT-DLC significantly enhanced the accuracy of matching keypoints (matches: $p < 0.01$) and reduced FN ($p < 0.001$), but increased FP ($p < 0.05$). No significant differences were found for ID switches (n.s) and RMSE (n.s) (Fig. 2E). When comparing SLEAP and vmT-LEAP, significant improvements were noted in Matches, FN, FP, and RMSE (SLEAP vs vmT-LEAP: Matches: $p < 0.01$, FN: $p < 0.01$, FP: $p < 0.001$, ID switches: n.s, RMSE: $p < 0.001$; Fig. 2E). In the comparison of vmTracking, vmT-DLC showed higher accuracy in Matches, FN, and ID switches than those in vmT-LEAP (vmT-DLC vs vmT-LEAP: Matches: $p < 0.001$, FN: $p < 0.001$, FP: n.s, ID switches: $p < 0.001$, RMSE: n.s, Fig. 2E).

In this study, the criterion for a Match was stringent, requiring both ID and body part to align correctly. Consequently, even if tracking was accurate after an ID switch, it was not considered a Match. Therefore, an evaluation was conducted by further classifying FN into cases where the correct body part was predicted but the ID was incorrect (body part match/ID mismatch FN), and cases where neither the correct body part nor the ID was predicted, including cases with missing predictions (body part mismatch FN) (Fig. S3). Repeated measures ANOVA revealed significant changes in both body part match/ID mismatch FN ($p < 0.01$) and body part mismatch FN ($p < 0.001$) (Greenhouse–Geisser-corrected, repeated-measures ANOVA, $n = 12$, Fig. S3). In pair-wise tests, no significant differences were observed

in body part match/ID mismatch FN (n.s). However, vmTracking generally maintained these at a low level (Fig. S3A). Conversely, vmTracking significantly reduced body part mismatch FN (maDLC vs vmT-DLC: $p < 0.001$, SLEAP vs vmT-LEAP: $p < 0.01$), with vmT-DLC showing a significant reduction compared to vmT-LEAP ($p < 0.001$, Fig. S3B).

The correlation between the ID match of virtual markers (Fig. S2) and the matching rates of vmTracking was depicted using a scatter plot (Fig. 2F), showing that both vmT-DLC ($r = 0.660$, $p < 0.05$) and vmT-LEAP ($r = 0.853$, $p < 0.001$) were positively correlated with virtual marker accuracy. Notably, vmT-LEAP achieved a matching rate (77.4%) comparable to virtual marker accuracy (73.9%), whereas vmT-DLC exceeded it with a matching rate of 91.4%. Therefore, vmTracking enhanced multi-animal tracking accuracy, particularly using DLC, in challenging contrast scenarios.

Tracking of black mice against white backgrounds

To validate the performance of vmTracking under different conditions, an experiment used three mice against a white background, focusing on scenarios classified as either occlusion-crowded (OC) or non-occlusion-crowded (nOC). This distinction allowed for a comprehensive assessment of vmTracking across different levels of complexity in the same environment, as well as differences in conditions. The experiment tackled the complexity of interpreting body part match/ID mismatch FN by evaluating markerless multi-animal pose tracking results after correcting scenes with retained ID switches of the entire set of keypoints, although it did not correct cases with temporary keypoint switches. We assigned six keypoints per mouse for markerless multi-animal pose tracking (Fig. 3, A and B left panel; Fig. 4, A and B left panel). Videos were then created to visualize two keypoints as virtual markers (Fig. 3A and 4A), and vmTracking was conducted (Fig. 3B right panel, Fig. 4B right panel, and Supplementary Video 1, 3, and 4). The results for maDLC and vmT-DLC were obtained after 11 iterations of retraining (with a final total of 3886 annotated frames) and five iterations of retraining (with a final total of 1,167 annotated frames), respectively. Additionally, to simplify the creation of virtual markers, we attempted maDLC by setting only the two points used as virtual markers as keypoints. However, this approach failed to stitch tracklets together, resulting in the inability to obtain coordinate data and, consequently, it was impossible to create virtual marker videos.

First, we confirmed that the behavioral syllables extracted by Keypoint-moseq (17) from the results of maDLC and vmT-DLC did not contradict the actual behaviors (Fig. 3C and 4C). While generally consistent with the actual behaviors (Fig. 3C of green and red syllables, 4C of all syllables), some behavioral syllables extracted by maDLC were identified as inconsistent in scenes with missing keypoints (Fig. 3C of blue syllables).

The study then performed a quantitative evaluation of the tracking results (Fig. 3D and 4D). In both OC and nOC scenes, vmTracking demonstrated significant improvements in matching accuracy (OC: $p < 0.001$, nOC: $p < 0.05$, Wilcoxon signed-rank test, OC: $n = 12$, nOC: $n = 8$) and FN (OC: $p < 0.001$, nOC:

$p < 0.05$), while maintaining consistent results in FP, ID switches, and RMSE (OC: n.s, nOC: n.s) (Fig. 3D and 4D). Remarkably, vmTracking not only enhanced the tracking accuracy in OC scenes but also showed improvements in nOC scenes where maDLC already exhibited a high matching rate (Fig. 4D of Matches).

Furthermore, we compared the relationship between the matching rate and the annotation frame counts for vmT-DLC and maDLC in OC scenes (Fig. 3E). In OC scenes, vmT-DLC achieved a tracking accuracy of 95.2% ($n = 12$) with a total of just 1,167 annotation frames, significantly outperforming maDLC, which was estimated to require 13,732 annotation frames based on a linear regression analysis to achieve comparable accuracy. This result highlights the efficiency of vmT-DLC in providing high-precision tracking with fewer training samples.

The correlation between the accuracy of the virtual markers and the matching rate of vmT-DLC was depicted using a scatter plot (Fig. 4E). This analysis revealed that when combining OC scenes ($n = 12$) and nOC scenes ($n = 8$), there was no significant positive correlation between the accuracy of virtual markers and the matching rate ($r = 0.434$, n.s, $n = 20$ in total). However, in OC scenes, vmT-DLC achieved a matching rate of 95.2%, surpassing the virtual marker accuracy of 89.5%. In nOC scenes, the matching rate was 97.5%, which was comparable to the virtual marker accuracy of 98.4%. Thus, vmT-DLC effectively enhanced the accuracy of multi-animal tracking in challenging scenarios, such as scenes with occlusion and crowding.

Tracking an object entering and exiting from a video frame

A validation study was conducted to simulate experiments where individuals enter and exit the video frame. Typically, maDLC does not detect markers of individuals not present in the frame, whereas saDLC outputs all markers regardless of individual presence. This setup explored the impact on tracking when individuals entered and exited the recording frame in vmT-DLC (Fig. S4). We manually removed and introduced one mouse at a time in conditions featuring three mice against a white background (Fig. S4A). During annotation, stray keypoints of mice removed from the frame were labeled on a remaining mouse. Accurate tracking was conducted with all three mice present, according to the virtual markers (Fig. S4A lower panel of 3 mice condition). However, when removing one mouse from the arena, the stray keypoints that lost their tracking target generally detected another remaining mouse, according to the annotations (Fig. S4A lower panel of 2 mice condition). Similarly, removing another mouse generally caused stray keypoints to detect the single remaining mouse (Fig. S4A lower panel of single-mouse conditions). When reintroducing the mice into the arena, tracking resumed with the correct IDs according to the virtual markers (Fig. S4A lower panel of transitions from one mouse to two- and three-mice conditions).

We calculated the distances between identical keypoints of the three mice in each frame and evaluated their distribution (Fig. S4, A and C). When defining less than 10 pixels between a stray keypoint and the

identical keypoint in another mouse as having overlap, we determined the frequency of such overlapping frames for each keypoint. The average percentage of frames in which stray keypoints overlapped with identical keypoints in another mouse was greater than 90% for the six keypoints ($p < 0.001$ for comparisons between all a and b pairs, two-way ANOVA with mixed design and Bonferroni correction for multiple comparisons, $n = 6$, Fig. S4B). This high overlap indicated that in vmT-DLC, stray keypoints that had lost their original tracking target aligned with the identical keypoints on any remaining mice. However, these overlapping keypoints were unnecessary in the finalized tracking results. To address this, we created a GUI-enabled code (https://doi.org/10.5281/zenodo.11337052) to remove overlapping keypoints (Fig. S4C). This approach effectively eliminated stray keypoints, ensuring tracking results were free from excess points (Fig. S4Da-c and Supplementary Video 1). Additionally, this code proved effective in maDLC for creating accurate virtual marker videos by removing unnecessary ID keypoints in the rare instances where keypoints from two IDs overlapped (Fig. S4Dd).

Verification of applicability to fish

The applicability of vmTracking to other animal species was validated by attempting to track a school of ten fish (Fig. 5A). For each fish, we tracked five equidistant points along the body axis, using the second and fourth keypoints as virtual markers (Fig. 5B). Virtual marker videos were created using SLEAP, and vmTracking was conducted using DLC (Supplementary Video 1 and 5). Accuracy was verified based on equidistant tracking along the body axis, using the variance of the distances between adjacent keypoints, where higher values reflect lower accuracy (Fig. 5C). Additionally, to detect schooling—a behavior characteristic of fish schools involving swimming in the same direction—the head direction of each fish was calculated, and their cosine similarity value (Cos sim) was measured based on the orientation of the two anterior keypoints (Fig. 5D).

The trajectories for keypoint 1 (head) in both maDLC and vmT-DLC were demonstrated (Fig. 5E). While maDLC showed intermittent trajectories, vmT-DLC consistently produced smooth trajectories. Fig. 5F displays snapshots of five exemplary scenes, along with their respective variance (σ^2) and Cos sim values, demonstrating that vmT-DLC more accurately tracked each fish. The comparison of median variance for each fish revealed significantly lower variance in vmT-DLC ($p < 0.01$, Wilcoxon signed-rank test, $n = 10$, Fig. 5G). Analysis also revealed significantly fewer frames with outlier variances in vmT-DLC when counting frames where each fish's variance values were outliers ($p < 0.01$, Fig. 5H). These results indicate that vmT-DLC more effectively detected keypoints at consistent intervals, essential for accurate tracking. In comparisons based on head direction, the frequency of head direction changes of more than 90° between frames (approximately 0.03 seconds) was significantly lower in vmT-DLC ($p < 0.01$, Fig. 5I). The time series variation of Cos sim between both maDLC and vmT-DLC appeared to show similar patterns; however, maDLC displayed more fine-scale variability (Fig. 5J), indicating more frequent changes in the similarity of head direction between frames.

Tracking of coordinated poses and movements among dancers

Seven human dancers were tracked (Fig. 6A), each assigned 19 keypoints, with those on the head, trunk, and both elbows and knees as virtual markers (Fig. 6B) in videos created using maDLC, and vmTracking was conducted using DLC (Supplementary Video 1 and 6). While long-term tracking revealed inaccuracies in keypoint predictions with both maDLC and vmT-DLC, maDLC occasionally switched the entire set of keypoints to a different dancer, whereas vmT-DLC did not exhibit such phenomena (Supplementary Video 6). The quantitative analysis focused on a scene where similar movements occurred, evaluating the accuracy of detecting the coordinated movements of the seven dancers during a two-second period. Fig. 6C shows representative examples of the temporal changes in the poses of three dancers tracked by maDLC and vmT-DLC. In maDLC, keypoints occasionally appeared in positions distant from their correct locations, considerably more than that in vmT-DLC. Actual images from these scenes were displayed to illustrate the real tracking images (Fig. 6D).

To quantitatively assess the similarity of the poses of the seven dancers, a normalized distance matrix for all keypoint pairs of each dancer was created and the root mean square deviation (RMSD) between the dancers was compared (lower RMSD indicates higher similarity) (Fig. 6E). vmT-DLC consistently showed RMSDs between 10 and 20, whereas maDLC sporadically displayed spikes in higher values, indicating lower pose similarity in some frames (Fig. 6F). Overall, vmT-DLC demonstrated significantly lower RMSD values ($p < 0.001$, Wilcoxon signed-rank test, $n = 41$, Fig. 6G). Similarly, the analysis evaluated scenes where the seven dancers maintained similar poses while stationary (Fig. S5). Even in these stationary scenes, maDLC showed keypoint switches, especially among overlapping dancers, whereas vmT-DLC achieved stable tracking (Fig. S5, A and B), and showed significantly lower average RMSD ($p < 0.001$, Fig. S5C).

Discussion

Despite advancements in markerless multi-animal pose tracking tools, overcoming occlusion and crowding remains a challenge. Studies using saDLC to simultaneously track one black C57BL/6N mouse and one white CD1 mouse have been reported (18), suggesting that tracking multiple animals with saDLC is feasible when there is a visual difference. Although attaching visually identifiable physical markers to animals could be a potential solution, physical markers may unpredictably affect behavior. In this study, we adopted a straightforward strategy of applying virtual markers to videos of markerless animals, allowing us to distinguish between animals without the need for physical markers. By tracking multiple animals with virtual markers using a single-animal pose tracking tool, we improved the accuracy of traditional multi-animal pose tracking in scenarios involving occlusion and crowding. In this section, we discuss the advantages and limitations of this method, namely vmTracking, as well as its implications for future research directions and its impact on this field.

vmTracking methodological enhancements and efficiency

vmTracking showed superior performance compared to traditional markerless multi-animal pose tracking methods, consistently improving accuracy across various conditions, including tracking three or five animals and managing occlusion and crowding. Notably, vmT-DLC improved keypoint matching rates in maDLC to approximately 90%, even in scenes where the rates were previously less than 50%. While vmT-LEAP did not perform as well as vmT-DLC, it still outperformed SLEAP. These results confirm vmTracking as a robust solution for addressing occlusion and crowding, likely contributing to improving the reliability of posture analysis tools like Keypoint-moseq (18).

The accuracy of vmTracking was positively correlated with virtual marker accuracy, suggesting that improving virtual marker accuracy could further enhance tracking precision. Notably, for vmT-DLC, even though only 70% to 80% of the virtual markers matched IDs correctly (with 20% to 30% of the expected virtual markers having ID mismatches or being missing), the accuracy of vmTracking still exceeded 90%. The accuracy criteria for virtual markers in this study were based on whether a label was attached to any part of an individual with a matching ID. Ensuring correct IDs for keypoints is more critical than precise keypoint matching when creating virtual markers, allowing some inaccuracies in keypoint matching. Additionally, some errors in ensuring correct IDs are also acceptable. Therefore, compared to the extensive effort required to manually correct every keypoint in every frame in traditional markerless multi-animal pose tracking, the effort needed for corrections when creating virtual markers is significantly less, making the process more efficient.

vmTracking also demonstrates exceptional efficiency in terms of the number of annotation frames required for training. In three-mice experiments, vmT-DLC achieved a keypoint matching rate of over 90% with approximately 1,200 annotated frames, even in scenarios with occlusion and crowding. In contrast, maDLC was estimated to require annotating more than ten times that number to achieve similar accuracy, which is impractical. Furthermore, while this estimation is based on simple regression, accuracy might plateau in reality, illustrating that vmT-DLC can achieve with minimal training what is extremely challenging to accomplish using maDLC.

Tracking when the number of subjects changes

In maDLC, when the number of target subjects in the video frames changes, keypoints of undetected subjects are not predicted. However, in saDLC, all keypoints are predicted regardless of the number of subjects present. This feature helps prevent the missed predictions that frequently occur in maDLC during occlusion and crowding, contributing to the enhanced accuracy of vmT-DLC. However, when the actual number of subjects changes, saDLC inevitably results in stray keypoints predicting alternative locations, which can potentially cause confusion in the tracking results. Additionally, there is a fundamental dilemma in how to manage these stray keypoints during refinement for retraining. To avoid the constant

effort of selecting and deleting these inevitably occurring stray keypoints, we attempted to apply them to the remaining mice during refinement. Importantly, this process involves labeling incorrect mice, which have virtual markers different from the correct ones that should be used for labeling the stray keypoints. When no target mouse is present, stray keypoints detect a provisional target, but when the correct mouse appears, it is prioritized for tracking. By annotating multiple-colored virtual markers, vmT-DLC can track different markers based on the situation. Consequently, vmT-DLC can handle stray keypoints without significantly compromising accuracy. Virtual marker colors are crucial for individual identification but are not absolute markers. This flexibility may contribute to vmT-DLC's ability to achieve high accuracy, even if the virtual markers are not strictly accurate.

Cross-species and setting applicability of vmTracking

The efficacy of vmTracking across various species and environments has been validated in both fish and humans. Tracking fish schools is considered a challenging task with maDLC (7), often resulting in noisy outputs characterized by frequent keypoint switches, which was observed in this study. However, implementing vmTracking mitigated these instabilities in maDLC tracking. For human tracking, vmTracking was applied to human dancing, but occasional failures in individual keypoint tracking were still observed during movement, suggesting the need for further improvements, such as refining the placement of virtual markers. Nonetheless, vmT-DLC did not exhibit keypoint set-level ID switches, where entire sets of keypoints switch rather than individual keypoints. Additionally, in low-movement scenes, tracking failures were rare despite occlusion and crowding. Thus, vmT-DLC demonstrated relatively stable accuracy in tracking the postures of multiple dancers compared to that in maDLC, which often encounters keypoint set-level ID switches and tracking errors even in low-motion scenes. This suggests vmTracking’s potential application in human scenarios such as sports analysis, including highly intermixed contact sports like soccer and basketball, where player interactions are frequent.

In human pose estimation (HPE), tools like OpenPose (19) and AlphaPose (20) are highly regarded for human tracking. However, HPE also faces challenges such as occlusion and crowding, with various improved approaches being explored to address these issues (21-23). Occlusion and crowding can lead to ambiguous or incorrect individual identification. Since vmTracking is particularly effective in addressing these issues, it presents a promising solution for the challenges in the HPE field, though there is still room for improvement in the accuracy of individual keypoints.

Optimizing virtual marker creation

Using one virtual marker per animal may suffice for individual identification, but we used two for mice and fish to ensure redundancy in case one keypoint is missing. Setting two virtual markers also helps clarify individual identification in scenarios where two animals overlap. While one might argue that more virtual markers could improve individual identification, it also increases the number of keypoints to

consider during the manual modification process. Moreover, since vmTracking relies on virtual markers, increasing their number could impact the tracking results. As our primary tracking targets are the animals themselves rather than the virtual markers, it is preferable to use fewer virtual markers to preserve the subjects' natural characteristics. However, for humans, larger subjects with independently moving limbs and heads , we assigned virtual markers to six points, including the trunk. Compared to mice and fish, tracking errors were more frequent for humans, suggesting the potential need for more virtual markers.

When creating virtual marker videos, tracking keypoints that will not be used as virtual markers is theoretically unnecessary. For example, in our human tracking experiments, tracking parts like wrists or ankles was unnecessary. Therefore, we attempted to create virtual markers with minimal keypoint tracking using maDLC in mouse experiments but failed to stitch tracklets, and thus could not obtain prediction results. This suggests that tracking more keypoints, even if they are not used as virtual markers, would likely yield better results. Although having many unnecessary keypoints complicates manual corrections in virtual marker creation, the process can be streamlined by deleting unnecessary keypoints and focusing solely on those intended for virtual markers. The code we developed to remove specified ranges of data from coordinate data files (h5 format) was initially intended to eliminate extraneous labels in vmT-DLC, but can also be used to remove unnecessary keypoints during virtual marker creation using maDLC.

Limitations

While vmTracking can be considered a method for high-accuracy tracking of multiple animals, it is important to recognize that there are several limitations that can make its application challenging.

First, vmTracking relies on virtual markers for individual identification. It can be nearly impossible to apply it in situations where there are too many targets to be identified by virtual markers alone. However, such situations are mostly encountered in natural environments with wild animal herds, and vmTracking should cause no issues in controlled environments like laboratories.

Second, from the perspective of preparing virtual marker videos, it is challenging to apply vmTracking to videos that are unmanageable with markerless multi-animal tracking. Specifically, it becomes difficult in cases of markerless multi-animal pose tracking where many frames fail to detect most targets, as seen in scene V of Fig. 5F in maDLC. However, aside from tools like maDLC and SLEAP that we used, various other markerless multi-animal tracking tools are already available (24-27), potentially allowing us to overcome this challenge.

Third, it is impossible to track scenarios where even the experimenter cannot determine the correct answer. If an animal is temporarily hidden during movement or if only a part of the body is obscured, it may be possible to generate plausible tracking results by estimating and adding virtual markers to the hidden parts (as in our mouse tracking studies). However, it is impossible to track situations where the

target is completely obscured by a large object, as this is essentially the same as not being present in the video frame. There is no way to track something that is not there.

In summary, the development and implementation of vmTracking in this study mark an advancement in the field of markerless multi-animal pose tracking. vmTracking has demonstrated exceptional accuracy and efficiency, particularly in addressing challenges such as occlusion and crowding, positioning it as a robust alternative to traditional tracking methods. Its reliance on virtual markers simplifies the tracking process, greatly reducing the need for manual annotation and training, making it more user-friendly and practical for both laboratory settings and broader applications. This efficiency allows for high accuracy with relatively few annotation frames, highlighting its potential for animal tracking in complex environments. Further research into the impact of the number, color, size, and position of virtual markers on tracking accuracy will open up possibilities for vmTracking. Overall, vmTracking is a useful tool in the study of animal behavior, ecology, and other related fields, providing an effective and efficient solution to some of the most persistent challenges in multi-animal tracking.

Materials and Methods

Multi-mouse dataset

Behavioral data were collected from C57BL/6J mice housed in groups of 2–3 per cage in an environment maintained at a temperature of 24–26°C and under a 12-hour light-dark cycle. These mice had unrestricted access to food and water. All behavioral recordings were performed during the light phase of the cycle. Cages and housing conditions were refreshed weekly. All experimental procedures received approval from the Institutional Animal Care and Use Committee of Doshisha University.

Data collection occurred in two different environments. In the first setting, five C57BL/6J mice freely roamed a rectangular open field with a black floor measuring 55 × 60 cm and walls 40-cm high. These recordings lasted approximately 30 minutes, captured at a resolution of 640 × 640 pixels and a frame rate of 30 fps. In the second environment, three C57BL/6J mice were placed in a circular open field with a white floor measuring 31 cm in diameter and transparent walls 17-cm high. Recordings in this setting were conducted for approximately 60 minutes at a resolution of 460 × 460 pixels and a frame rate of 30 fps. For this second setting, two different recording sessions involving three mice were prepared: one was a continuous recording lasting 60 minutes, and another involved manipulating the number of mice in the video frames by introducing or removing them from the arena at approximately 12-minute intervals. In the latter session, temporary IDs were assigned to three mice, with the sequence as follows: three mice for the first 12 minutes, two mice (IDs 1 and 2) for the next 12 minutes, one mouse (ID 1 only) for the subsequent 12 minutes, two mice (IDs 1 and 3) for the next 12 minutes, and finally three mice for the last 12 minutes, as depicted in Fig. S4A.

Multi-fish dataset:

A previously published video featured 10 fish swimming in a rectangular aquarium, recorded at 30 fps (28). For our analysis, the video was scaled down to 1228 × 1030 pixels. The video is available at: https://drive.google.com/drive/folders/1Xn3IY46t7DkG2bntbO6sL3oElRPeQA1d.

Multi-human dance dataset:

We used a publicly available dataset that featured a video of seven dancers wearing uniforms, recorded at 20 fps (29). The video is available at: https://drive.google.com/drive/folders/1ASZCFpPEfSOJRktR8qQ_ZoT9nZR0hOea.

DLC as markerless multi-animal pose tracking

In this study, we utilized DLC version 2.2.3 for the maDLC project. For the mouse tracking experiments, 50 frames were extracted from each video, and six keypoints (along the midline from the nose to the base of the tail) were labeled for scenarios involving either five or three mice. Similarly, in the fish tracking experiments, 50 frames per video were extracted and five keypoints were equidistantly labeled along the body axis, with the first keypoint positioned at the head end. Skeletons were created for all keypoint pairs in both mouse and fish tracking experiments. In the dance tracking experiments, 20 frames were extracted, and a total of 19 keypoints, including the head, neck, trunk, shoulders, elbows, wrists, hands, leg bases, knees, ankles, and feet, were labeled for annotation.

The DLCRnet, known for its high performance in tracking multiple subjects (7), was employed as the training network across all tracking experiments using maDLC. After the initial training, performance evaluation and video analysis were conducted. Based on the results, outlier frames were identified, annotated, and subjected to further training. For the mouse tracking experiments, this refinement and training process was repeated five times, resulting in annotations for a total of 1,097 frames for five mice and eleven times for a total of 3,886 frames for three mice. The fish tracking experiments underwent five rounds of refinement and training, resulting in annotations for a total of 811 frames. Similarly, the dance tracking experiments were refined and retrained five times, culminating in annotations for a total of 112 frames. Coordinate data were obtained after each training session and utilized to assess the tracking accuracy.

Creating virtual marker video using DLC

Virtual marker videos for vmTracking of three mice were created based on the results from iteration-6 of maDLC, which included annotations of 2820 frames. These videos ensured that all three mice were continuously present within the video frames. For videos where the number of mice present in the video frames was varied by removing or reintroducing mice into the arena, the videos were analyzed using iteration-11 of the training data. maDLC stitches together tracklets, which are fragments of motion

trajectories obtained from video analysis, to form continuous behavioral trajectories for each animal (7). However, the generation of tracklets can be interrupted by occlusion, leading to failures in stitching and resulting in ID switches. These ID switches, including partial ID switches at the six keypoints for each mouse, were manually corrected. However, these corrections were not always perfect. Specifically, corrections were not necessarily made in cases where there was a temporary switch in ID or when the predicted location of the keypoints was off, but the predicted ID was correct. The corrected data were used to create tracking videos, which were then designated as virtual marker videos. Only keypoints 2 and 4 of each mouse were used as virtual markers. The keypoints in these videos were displayed using the matplotlib colormap 'gray', with a marker size set to 2, and no skeletons were outputted. Moreover, to further simplify the creation of virtual marker videos, we also performed maDLC by tracking only the two keypoints that were intended to be designated as virtual markers.

For the vmTracking of dancers, virtual marker videos were created based on results from iteration-5 of maDLC. In these videos, six of the 19 keypoints—specifically the head, trunk, and both elbows and knees—were utilized as virtual markers. The videos were produced using the matplotlib colormap 'rainbow', with a marker size of 5, and no skeletons were outputted.

vmTracking using DLC

The saDLC project was established to analyze virtual marker videos. In the mouse tracking experiments, the number of keypoints per mouse was set at six, and the total number of keypoints to be labeled was determined by multiplying this number by the number of mice tracked (i.e., a total of 30 keypoints for five mice and 18 keypoints for three mice). Each mouse was allocated six keypoints (along the midline from the nose to the base of the tail), and skeletons were constructed for all keypoint pairs, tracking similar parts as in maDLC. For annotation, 50 frames were extracted, similar to maDLC, and annotations were made by assigning consistent IDs across frames, based on the color of the virtual markers. In scenarios where the number of mice in the video frames changed, stray keypoints labeled the identical parts of the remaining mice. In frames where only two mice were present in the arena, labeling was done such that IDs 2 and 3 overlapped, and if only one mouse was present, all labels were assigned to that single mouse.

In the fish tracking experiments, five keypoints per fish were set, with the total number of keypoints determined by multiplying this number by the number of fish tracked (50 keypoints in total). Similarly, each fish was allocated five keypoints, and skeletons were constructed for all keypoint pairs, again tracking similar parts as in maDLC. Fifty frames were extracted for annotation, with consistent IDs assigned across frames, based on the color of the virtual markers.

In the dance tracking experiments, each dancer was assigned 19 keypoints, with the total number of keypoints set by multiplying this by the number of dancers tracked (133 keypoints in total). Nineteen

keypoints were allocated per dancer, tracking similar parts as in maDLC. For annotation, 20 frames were extracted, and IDs were consistently assigned across frames, based on the color of the virtual markers.

The training network selected was EfficientNet_b0, known for its superior performance and lower computational costs compared to ResNet or MobileNet (30). After the training phase, performance evaluation and video analysis were conducted. Based on these results, outlier frames were identified, annotated, and subjected to further training. These refinement and training processes were repeated five times for the five-mice tracking experiments, resulting in a total of 1,196 annotated frames. For the three-mice tracking experiments, these processes were also repeated five times, accumulating a total of 1,167 annotated frames. In the fish tracking experiments, the process was repeated five times, resulting in a total of 885 annotated frames. Similarly, in the dance tracking experiments, the process was repeated five times, yielding a total of 294 annotated frames. Finally, the predicted results were output to the original markerless videos to obtain tracking videos with the virtual markers removed.

Social LEAP as markerless multi-animal pose tracking

In this study, SLEAP version 1.3.3 was employed. The annotation data used in maDLC were imported, and training and prediction were conducted using SLEAP. The training type employed a “multi-animal top-down” approach, and maximum instances were set to match the number of animals being analyzed. In prediction, a simple tracker method was selected, and the maximum number of tracks was set equal to the number of animals analyzed. The obtained coordinate data were used to evaluate tracking accuracy.

Creating virtual marker video using SLEAP

Virtual markers for the five-animal tracking experiment were created based on the SLEAP results. From the predicted results of markerless multi-animal pose tracking, all predicted keypoints except for keypoints 2 and 4 which served as virtual markers, were deleted. Subsequently, corrections were made to the ID switches and keypoint positions. Frames in which only one mouse was predicted were extracted, annotations for the four unpredicted mice were added, followed by retraining and re-prediction. Additionally, frames in which only two mice were predicted were extracted. Random frames were selected, annotations for the three unpredicted mice were added, and further training and predictions were performed. Temporary ID switches or cases where the predicted ID was correct despite a misalignment in the keypoint prediction position were not necessarily corrected. At this stage, labels for the unpredicted mice were not added. Using these corrected data, tracking videos were created and designated as virtual marker videos. During video creation, the marker size was set to 2, and skeletons were not output.

Virtual markers for the fish tracking experiment were created following the SLEAP results. After making corrections to the ID switches and keypoint positions based on the predicted results of markerless multi-animal pose tracking, all predicted keypoints except for keypoints 2 and 4, which served as virtual markers, were deleted to create the videos. The marker size was set to 2, and skeletons were not output.

vmTracking using SLEAP's single-animal method

The annotation data used in vmT-DLC were imported for training and prediction. The training type employed was the single-animal method of SLEAP, which is equivalent to LEAP. For prediction, a simple tracker method was selected, and the maximum number of tracks was limited to one. The obtained coordinate data were utilized to assess tracking accuracy. In addition, tracking videos displaying the predicted results were created.

Mice tracking performance evaluation

The accuracy of multi-mouse tracking was evaluated based on a manually created ground truth for sampling sets of either 150 or 300 frames. This evaluation focused on tracking keypoints. To compare predicted data with ground truth, each predicted mouse ID was paired with the most matching ground truth ID. For this purpose, distances for all combinations of predicted data and ground truth IDs were calculated, and the Hungarian algorithm was used to determine the most suitable combinations. The evaluation metrics included Match, FN, FP, ID switch, and RMSE. Match, synonymous with the percentage of correct keypoints, was defined as the number of instances in which the predicted point is detected within a predefined threshold (10 pixels) from the ground truth point (Fig. S1, A and B). This required a match in both body part and ID with the ground truth. Cases not qualifying as Matches were assessed as FN. FN represents keypoints where correct predictions were not made, defined as either predictions within the threshold for an non-corresponding ground truth point or missing points (Fig. S1, A and B). Even in the former case, if the prediction is within the threshold for the corresponding ground truth point, it is prioritized as a Match. Furthermore, FN were classified into two categories: one where the ID differed but the predicted part matched (body part match/ID mismatch FN), and another where the predicted part was incorrect irrespective of the ID (body part mismatch FN) (Fig. S1B). FP represents keypoints that predict non-existent points, defined as points detected outside the threshold from all ground truth points (Fig. S1, A and B). An ID switch is defined as an instance in which predictions are allocated within the threshold to any ground truth point of a different mouse across frames (Fig. S1, A and B). RMSE was calculated based on pairs of corresponding predicted data and ground truth data.

Accuracy of virtual markers

The accuracy of the virtual markers was defined based on the identity matching rate for keypoints 2 and 4 (points becoming virtual markers) after correcting markerless multi-animal pose tracking results. In other words, an ID match was considered when the indicated ID of the virtual marker matched any ground truth for the same mouse within a threshold (10 pixels) (Fig. S2). The accuracy of the virtual markers was calculated as the ratio of virtual markers with matching IDs. If keypoints designated to

become virtual markers were missing, the corresponding virtual markers would also be absent. These missing virtual markers were treated as mismatches.

Keypoint-moseq

Keypoint-moseq (18) was used to extract behavioral pattern features (syllables) from coordinate data obtained through maDLC and vmT-DLC. The fitting process involved 50 iterations for the AR-HMM and an additional 500 iterations for the full model, with the kappa value set at 1e3 (1000). The extracted syllables were subsequently checked for consistency with the actual movements of the mice.

Evaluation of stray keypoints

In experiments where mice were introduced or removed from the arena, we calculated the distances between identical keypoints for each ID to verify that we could track the remaining mice as temporary targets using stray keypoints that had lost their original targets. Keypoints were considered to overlap if the distance between them was within 10 pixels, and the frequency of such frames was assessed. Since overlapping stray keypoints were unnecessary in the cleaned and finalized dataset, we created a GUI-enabled code (https://doi.org/10.5281/zenodo.11337052) to remove these redundant keypoints (Fig. S4C). This cleaned and finalized dataset was then used to create the videos.

Fish tracking evaluation

Based on setting equidistant keypoints along the body axis, it was assumed that if the distances between adjacent keypoints were not uniform, the tracking was not accurate. The focus of the evaluation was on whether the tracking accurately maintained the equidistant five keypoints. To this end, the variance of distances between adjacent keypoints was calculated as a measure of tracking accuracy, with greater variance interpreted as lower accuracy. The variance for each fish per frame was calculated, the median variance identified, and the frequency of outliers assessed. Outlier thresholds were determined using the interquartile range method based on integrated data from maDLC and vmTracking. For each condition, these thresholds were applied, and the number of frames with variance considered outliers for each fish was counted. Furthermore, it is rare for the direction of the head to change abruptly between frames, and it was assumed that such changes occur when the predictions of keypoints are not synchronized. Therefore, the angle of head direction was calculated using the two anterior keypoints for each fish, with a focus on the synchronization of these predictions. The difference in angles between frames was measured, and the changes exceeding 90 degrees were counted as abnormal directional changes. More such frames were interpreted as lower tracking accuracy.

Dance tracking evaluation

In this study, we evaluated the detection of similar poses during two seconds of similar movements and two seconds of static poses among seven dancers. For each dancer, we calculated distances between all keypoints to create a standardized distance matrix. The RMSD of the standardized distance matrices between dancers was then computed as a measure of similarity; the more similar the poses, the smaller the RMSD. We observed the variation in the average RMSD values for all dancer pairs per frame to determine the similarity for each frame, and averaged these over all frames to measure the overall similarity for the scene.

Statistical analysis

In the tracking experiment with five mice, 12 OC scenes where keypoints from different mice intersected were sampled from four independent videos. The tracking accuracy of maDLC, vmT-DLC, SLEAP, and vmT-LEAP was evaluated using repeated-measures ANOVA and multiple comparisons with Bonferroni correction, based on the previously mentioned metrics. In the tracking experiment with three mice, 12 OC and eight nOC scenes were sampled from four independent videos. The tracking accuracies of maDLC and vmT-DLC were compared using the Wilcoxon signed-rank test. Additionally, a correlation analysis was conducted between the matching rate of vmTracking and the accuracy of the virtual markers based on the ID matching rate. Furthermore, for the OC scenes in the three-mouse tracking experiment, a regression analysis using the least squares method was conducted to estimate the number of annotation frames required for maDLC to reach the accuracy level of vmT-DLC based on the matched keypoint rate.

For the evaluation of stray keypoints, the frequency of frames where the distance between identical keypoints was within 10 pixels was calculated for each of the six keypoints, and the average for each ID pair (ID1-ID2, ID1-ID3, ID2-ID3) was compared using a two-way mixed ANOVA with Bonferroni-corrected post hoc tests.

In the fish tracking experiment, the median variance from each fish and the frequency of frames where the variance indicated outliers, as well as the frequency of frames showing an abnormal change in head direction greater than 90 degrees, were compared using the Wilcoxon signed-rank test.

In the dance tracking experiment, the average RMSD value over the entire analyzed period was compared using paired t-tests.

These analyses utilized MATLAB functions ranova, anovan, multcompare, signrank, ttest, corrcoef, and polyfit. Statistical significance was set at $p < 0.05$.

References

1. A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, M. Bethge, DeepLabCut: Markerless Pose Estimation of User-defined Body Parts with Deep Learning. *Nature Neuroscience* **21**, 1281–1289 (2018).
2. T. D. Pereira, D. E. Aldarondo, L. Willmore, M. Kislin, S. S. -h. Wang, M. Murthy, J. W. Shaevitz, Fast animal pose estimation using deep neural networks. *Nature Methods* **16**, 117–125 (2018).
3. K. C. Newton, D. Kacev, S. R. O. Nilsson, A. L. Saettele, S. A. Golden, L. Sheets, Lateral line ablation by ototoxic compounds results in distinct rheotaxis profiles in larval zebrafish. *Communications Biology* **6** (2023).
4. M. Bogachev, A. Sinitca, K. Grigarevichius, N. Pyko, A. Lyanova, M. Tsygankova, E. Davletshin, K. Petrov, T. Ageeva, S. Pyko, D. Kaplun, A. Kayumov, Y. Mukhamedshina, Video-based marker-free tracking and multi-scale analysis of mouse locomotor activity and behavioral aspects in an open field arena: A perspective approach to the quantification of complex gait disturbances associated with Alzheimer's disease. *Frontiers in Neuroinformatics* **17** (2023).
5. K. Yamamoto, Y. Hasegawa, T. Suzuki, H. Suzuki, H. Tanabe, K. Fujii, Extracting proficiency differences and individual characteristics in golfers' swing using single-video markerless motion analysis. *Frontiers in Sports and Active Living* **5** (2023).
6. A. Syeda, L. Zhong, R. Tung, W. Long, M. Pachitariu, C. Stringer, Facemap: a framework for modeling neural activity based on orofacial tracking. *Nature Neuroscience*, doi: 10.1038/s41593-023-01490-6 (2023).
7. J. Lauer, M. Zhou, S. Ye, W. Menegas, S. Schneider, T. Nath, M. M. Rahman, V. Di Santo, D. Soberanes, G. Feng, V. N. Murthy, G. Lauder, C. Dulac, M. W. Mathis, A. Mathis, Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods* **19**, 496–504 (2022).
8. T. D. Pereira, N. Tabris, A. Matsliah, D. M. Turner, J. Li, S. Ravindranath, E. S. Papadoyannis, E. Normand, D. S. Deutsch, Z. Y. Wang, G. C. McKenzie-Smith, C. C. Mitelut, M. D. Castro, J. D'Uva, M. Kislin, D. H. Sanes, S. D. Kocher, S. S. -h. Wang, A. L. Falkner, J. W. Shaevitz, M. Murthy, SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods* **19**, 486–495 (2022).
9. S. Gül, S. Bosse, D. Podborski, T. Schierl, C. Hellge, Kalman filter-based head motion prediction for cloud-based mixed reality. *Proceedings of the 28th ACM International Conference on Multimedia*, 3632-3641 (2020)
10. B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking. *arXiv (Cornell University)*, (2018).
11. P. Dendorfer, A. Os□Ep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, L. Leal-Taixé, MOTChallenge: a benchmark for Single-Camera Multiple Target Tracking. *International Journal of Computer Vision* **129**, 845–881 (2020).

12. S. Han, H. Wang, E. Yu, Z. Hu, ORT: Occlusion-robust for multi-object tracking. *Fundamental Research*, (2023).
13. W. S. Kim, M. I. Khot, H.-M. Woo, S. Hong, D.-H. Baek, T. Maisey, B. Daniels, P. L. Coletta, B.-J. Yoon, D. G. Jayne, S. I. Park, AI-enabled, implantable, multichannel wireless telemetry for photodynamic therapy. *Nature Communications* **13** (2022).
14. D. J. Terstege, M. Dawson, N. F. Jamani, M. Tsutsui, J. R. Epp, D. Sargin, Protocol for the integration of fiber photometry and social behavior in rodent models. *STAR Protocols* **4**, 102689 (2023).
15. E. C. Wright, P. X. Luo, H. C. Zakharenkov, A. S. Godoy, A. A. Lake, Z. D. Prince, S. Sekar, H. I. Culkin, A. V. Ramirez, T. Dwyer, A. Kapoor, C. Corbett, L. Tian, A. S. Fox, B. C. Trainor, Sexual differentiation of neural mechanisms of stress sensitivity during puberty. *Proceedings of the National Academy of Sciences of the United States of America* **120** (2023).
16. L. Haalck, M. Mangan, A. Wystrach, L. Clement, B. Webb, B. Risse, CATER: Combined Animal Tracking & Environment Reconstruction. *Science Advances* **9** (2023).
17. C. Weinreb, J. Pearl, S. Lin, M. A. M. Osman, L. Zhang, S. Annapragada, E. Conlin, R. Hoffman, S. Makowska, W. F. Gillis, M. Jay, S. Ye, A. Mathis, M. W. Mathis, T. Pereira, S. W. Linderman, S. R. Datta, Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *bioRxiv (Cold Spring Harbor Laboratory)*, (2023).
18. L. Miranda, J. Bordes, M. Reinhardt, S. Narayan, J. Hartmann, E. L. Newman, L. M. Brix, V. D. Lotte, C. Engelhardt, L. Dillmann, S. Mitra, K. J. Ressler, B. Pütz, F. Agakov, B. Müller-Myhsok, M. Schmidt V., Automatically annotated motion tracking identifies a distinct social behavioral profile following chronic social defeat stress. *Nature Communications* **14** (2023).
19. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D pose Estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 172–186 (2021).
20. H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *arXiv (Cornell University)*, (2022).
21. L. Fortini, M. Leonori, J. M. Gandarias, D. M. Elena, A. Ajoudani, Markerless 3D human pose tracking through multiple cameras and AI: Enabling high accuracy, robustness, and real-time performance. *arXiv (Cornell University)*, (2023).
22. X. Chen, J. Zhang, K. Wang, P. Wei, L. Lin, Multi-Person 3D Pose esitmation with occlusion reasoning. *IEEE Transactions on Multimedia*, 1–13 (2024).
23. A. Glandon, L. Vidyaratne, N. K. Dhar, J. O. Familoni, N. Sadeghzadehyazdi, S. T. Acton, K. M. Iftekharuddin, 3D far-field Lidar sensing and computational modeling for human identification. *Applied Optics* **63**, C15 (2023).

24. F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. H. Heras, G. G. De Polavieja, idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nature Methods* **16**, 179–182 (2019).
25. T. Walter, I. D. Couzin, TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife* **10** (2021).
26. Z. Chen, R. Zhang, H.-S. Fang, Y. E. Zhang, A. Bal, H. Zhou, R. R. Rock, N. Padilla-Coreano, L. R. Keyes, H. Zhu, Y.-L. Li, T. Komiyama, K. M. Tye, C. Lu, AlphaTracker: a multi-animal tracking and behavioral analysis tool. *Frontiers in Behavioral Neuroscience* **17** (2023).
27. V. A. Le, T.-L. Sterley, N. Cheng, J. S. Bains, K. Murari, Markerless mouse tracking for social experiments. *ENeuro* **11**, ENEURO.0154-22.2023 (2024).
28. T. Niizato, K. Sakamoto, Y.-I. Mototake, H. Murakami, T. Tomaru, Information structure of heterogeneous criticality in a fish school. *bioRxiv (Cold Spring Harbor Laboratory)*, (2024).
29. P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, P. Luo, DanceTrack: Multi-Object tracking in uniform appearance and diverse motion, *arXiv.org* (2021).
30. M. Tan, Q. Le V., EfficientNet: Rethinking model scaling for convolutional neural networks, *arXiv.org* (2019).

Acknowledgements

Funding

This work was supported by Kakenhi grants from the Japan Society for the Promotion of Science (21H04247 and 24K15711 to H.A., and 21H05296 and 23H00502 to S.T.) and by the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Agency (JPMJCR23P2 to S.T.).

Author contributions

H. A. and S. T. conceived the study. H. A. acquired the behavioral data of the mice. H. A. performed the animal tracking. H. A. performed the formal analyses. H. A. and S. T. prepared the manuscript.

Competing interests

The authors declare that they have no competing interests.

Data and code availability

The results of this paper can be reproduced using downloadable data and accompanying code, which are available from the Zenodo repository (https://doi.org/10.5281/zenodo.11337052).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this study, the authors used ChatGPT to improve readability and language. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Figures

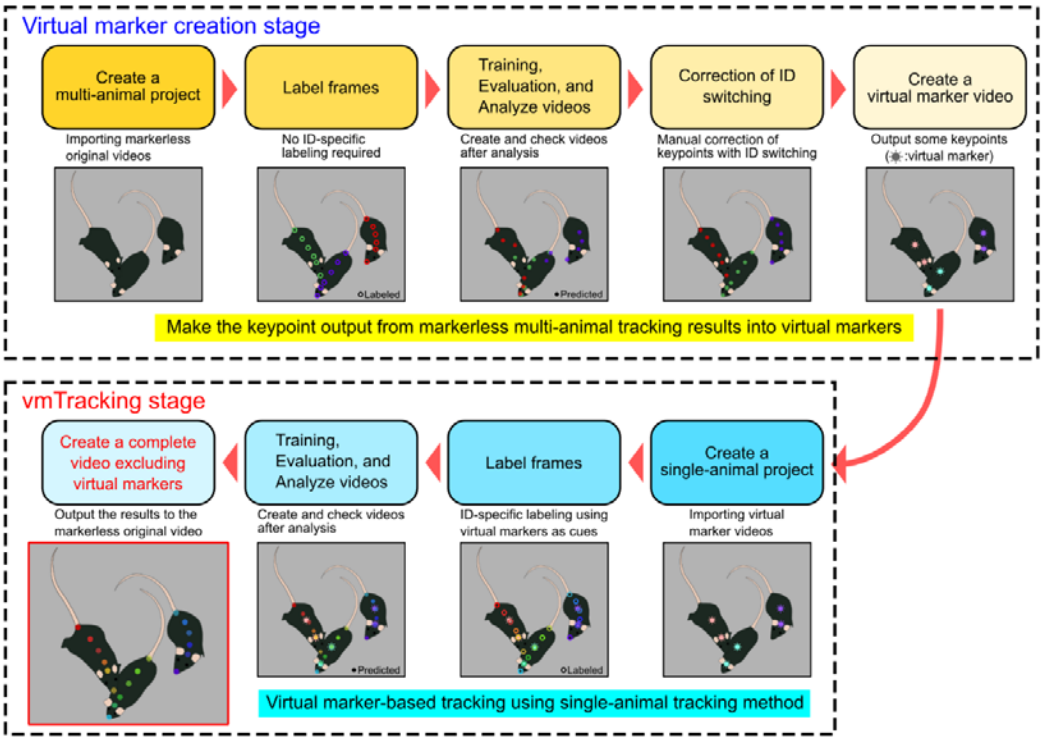

Fig. 1. vmTracking workflow.

vmTracking comprises two distinct phases: creating virtual marker videos and tracking these markers. Initially, a traditional markerless multi-animal pose tracking tool is employed to create the virtual marker videos. The initial tracking results are manually corrected to rectify ID switches and keypoint inaccuracies. Subsequently, a video that displays these refined keypoints is created. These keypoints are then designated as virtual markers for use in the next phase of tracking. During this phase, the virtual markers act as unique identifiers for annotating each animal individually. Ultimately, the refined tracking data are overlaid on the original markerless footage, resulting in a composite multi-animal tracking video that does not visibly include the virtual markers.

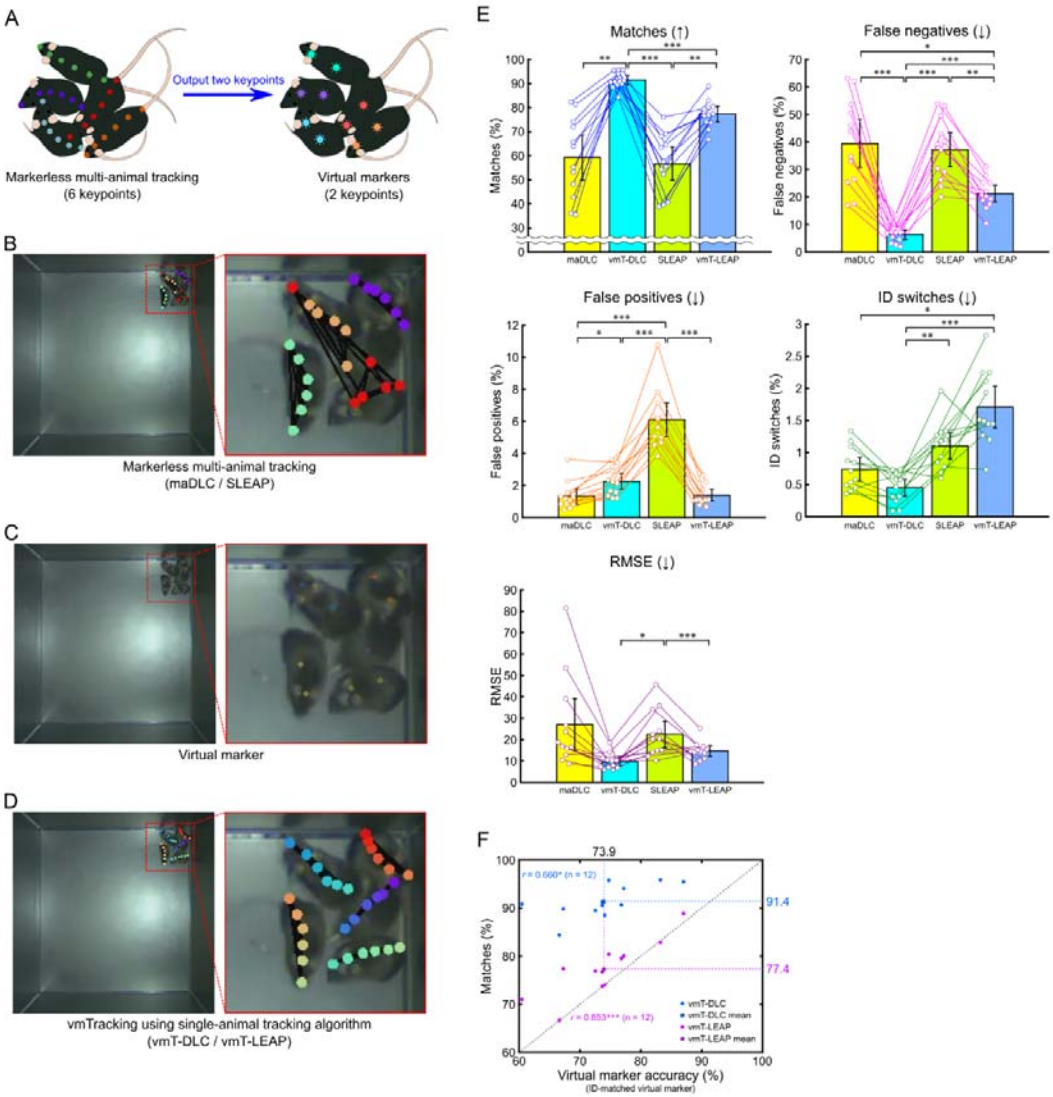

Fig. 2. Comparison of markerless multi-animal pose tracking and vmTracking in five mice tracking.

(A) Schematic of virtual marker creation. Six keypoints per mouse were tracked; two designated as virtual markers. (B–D) Snapshots of markerless multi-animal pose tracking in occlusion-crowded scenes (B), virtual markers (C), and vmTracking (D). Red lines on the right panel represent enlarged view of the red-lined box in the left panel. (E) For 12 sampled occlusion-crowded scenes (1,800 frames in total), the matching rate (Matches), false negative rate (False negatives), false positive rate (False positives), ID switch rate (ID switches), and root mean square error (RMSE) were calculated by comparing them with manually created ground truth for multi-animal DeepLabCut (maDLC), vmTracking with single-animal DeepLabCut (vmT-DLC), Social LEAP (SLEAP), and vmTracking with LEAP (vmT-LEAP). Arrows

indicate whether higher or lower values are better. Data represent mean ± 95% confidence interval (n = 12 scenes). Statistical tests used repeated-measures ANOVA and Bonferroni correction. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (F) Scatter plots illustrate the accuracy of virtual markers based on ID matching rate and corresponding matching rates in vmTracking for vmT-DLC and vmT-LEAP. Filled circles represent individual scenes, and the stars signify their average. Correlation analysis was conducted (n = 12 scenes). *: $p < 0.05$, ***: $p < 0.001$.

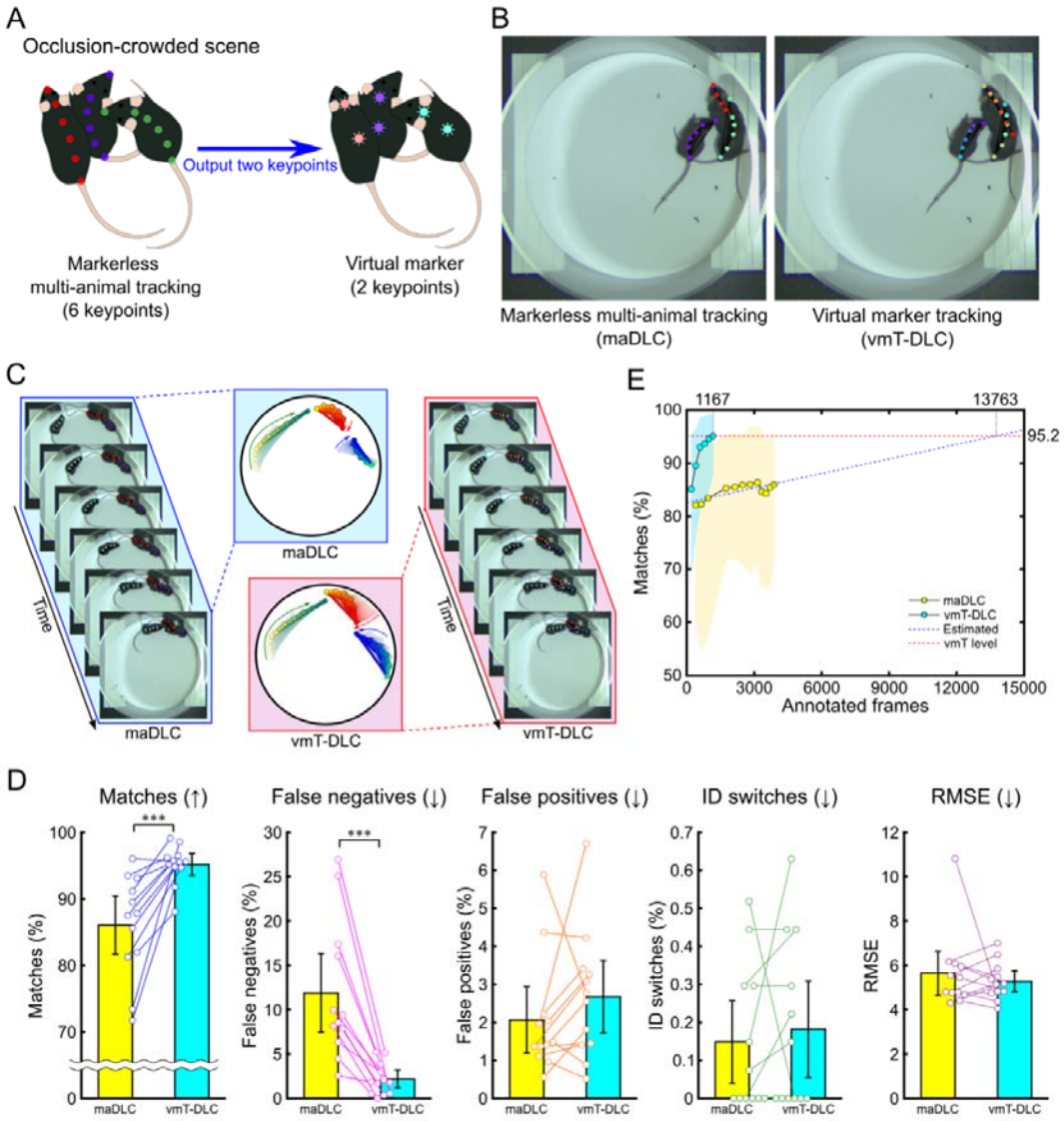

Fig. 3. Comparison of markerless multi-animal pose tracking and vmTracking in an occlusion-crowded scene of three mice.

(A) Six keypoints per mouse were tracked, with two designated as virtual markers. (B) Examples of markerless multi-animal pose tracking (left) and vmTracking (right) in an occlusion-crowded (OC) scene. (C) Posture patterns (Syllables) modeled by Keypoint-moseq for maDLC and vmT-DLC in an OC scene, with images taken every six frames showing movement direction. (D) For 12 OC scenes (1,950 frames in total), keypoints were compared with manually created ground truths, calculating matching rates (Matches), false negative rates (False negatives), false positive rates (False positives), ID switch rates (ID switches), and root mean square errors (RMSE). Multi-animal DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (vmT-DLC) were compared. Arrows indicate better values. Data are

shown as mean ± 95% confidence interval (n = 12 scenes). Statistical analysis used the Wilcoxon signed-rank test. ***: $p < 0.001$. (E) Estimating required annotation frames for markerless tracking to match vmTracking accuracy. The relationship between annotation frames and average matching rate for maDLC and vmT-DLC in OC scenes (three mice) was demonstrated. Regression analysis estimated the annotation frames required to reach the accuracy level of vmT-DLC based on the matching rate. Yellow markers represent maDLC, teal markers represent vmT-DLC. Yellow and teal shadows indicate the range of minimum and maximum values, respectively. The blue dotted line represents the regression line obtained from a simple linear regression analysis of maDLC performance, and the red dotted line indicates vmT-DLC's maximum average matching rate (95.1%, n = 12 scenes).

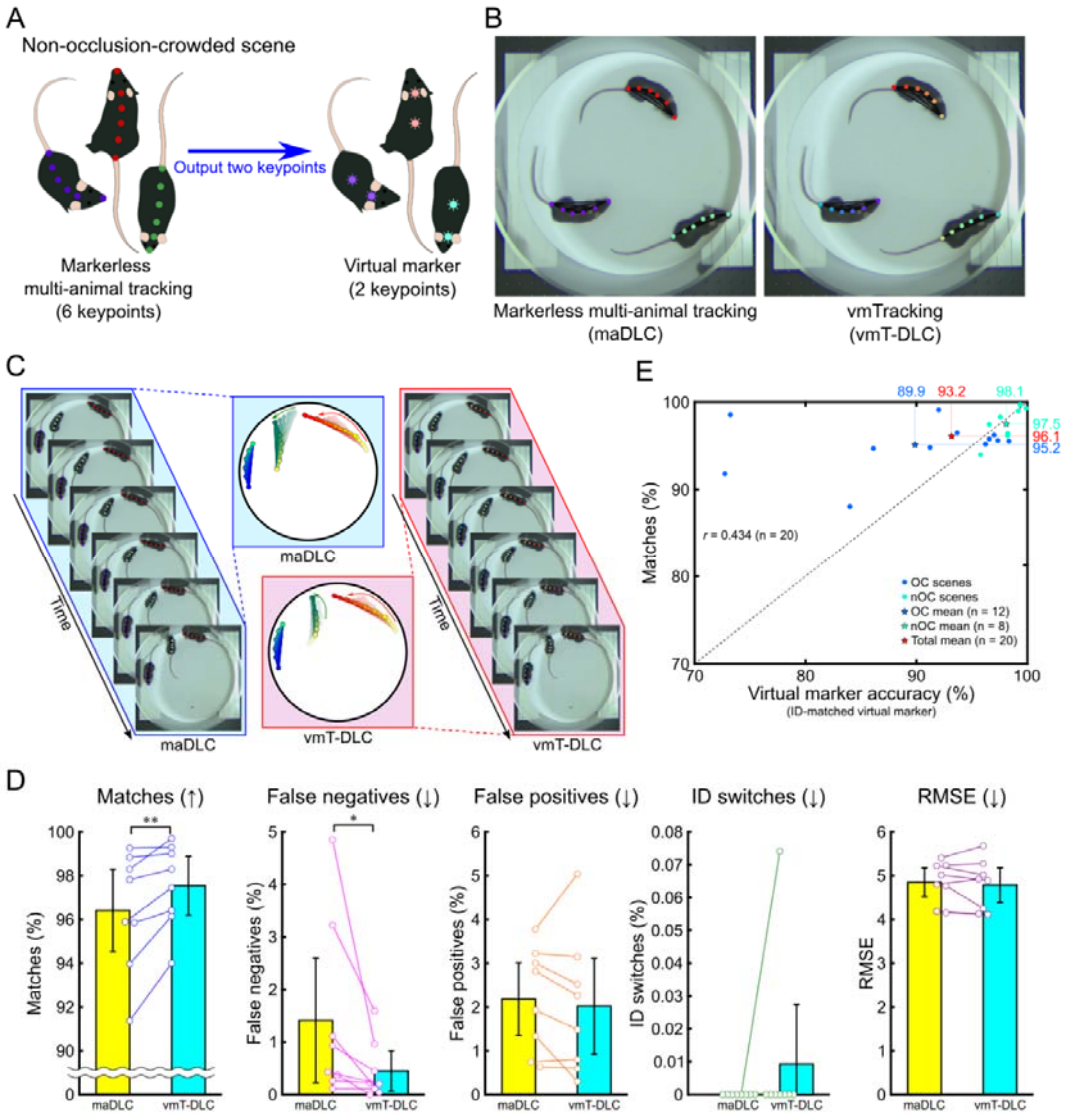

Fig. 4. Comparison of markerless multi-animal pose tracking and vmTracking in a non-occlusion-crowded scene of three mice.

(A) Schematic of virtual marker creation. Six keypoints per mouse were tracked, with two designated as virtual markers. (B) Snapshot examples of markerless multi-animal pose tracking (left) and vmTracking (right) in a non-occlusion-crowded (nOC) scene. (C) Posture patterns (Syllables) modeled by Keypoint-moseq for maDLC and vmT-DLC in an nOC scene. Images were taken every six frames, showing each mouse's Syllable corresponding to that scene. Arrows indicate general direction of movement. (D) For eight sampled nOC scenes (1,350 frames in total), keypoints were compared with manually created ground truths, to obtain matching rates (Matches), false negative rates (False negatives), false positive rates (False positives), ID switch rates (ID switches), and root mean square errors (RMSE).

Arrows indicate better values in DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (vmT-DLC) comparison. Data are shown as mean ± 95% confidence interval (n = 8 scenes). Statistical analysis was performed using the Wilcoxon signed-rank test. *: $p < 0.05$, **: $p < 0.01$. (E) Correlation between the accuracy of virtual markers and the matching rate of vmT-DLC when combining occlusion-crowded (OC) and nOC scenes. The scatter plot shows the accuracy of the virtual markers based on the ID matching rate and the corresponding concordance rate of vmT-DLC. Scenes were differentiated by color into OC (n = 12) and nOC (n = 8), for a total of 20 scenes. The filled circles represent the values for each sampling scene, and the stars indicate their average.

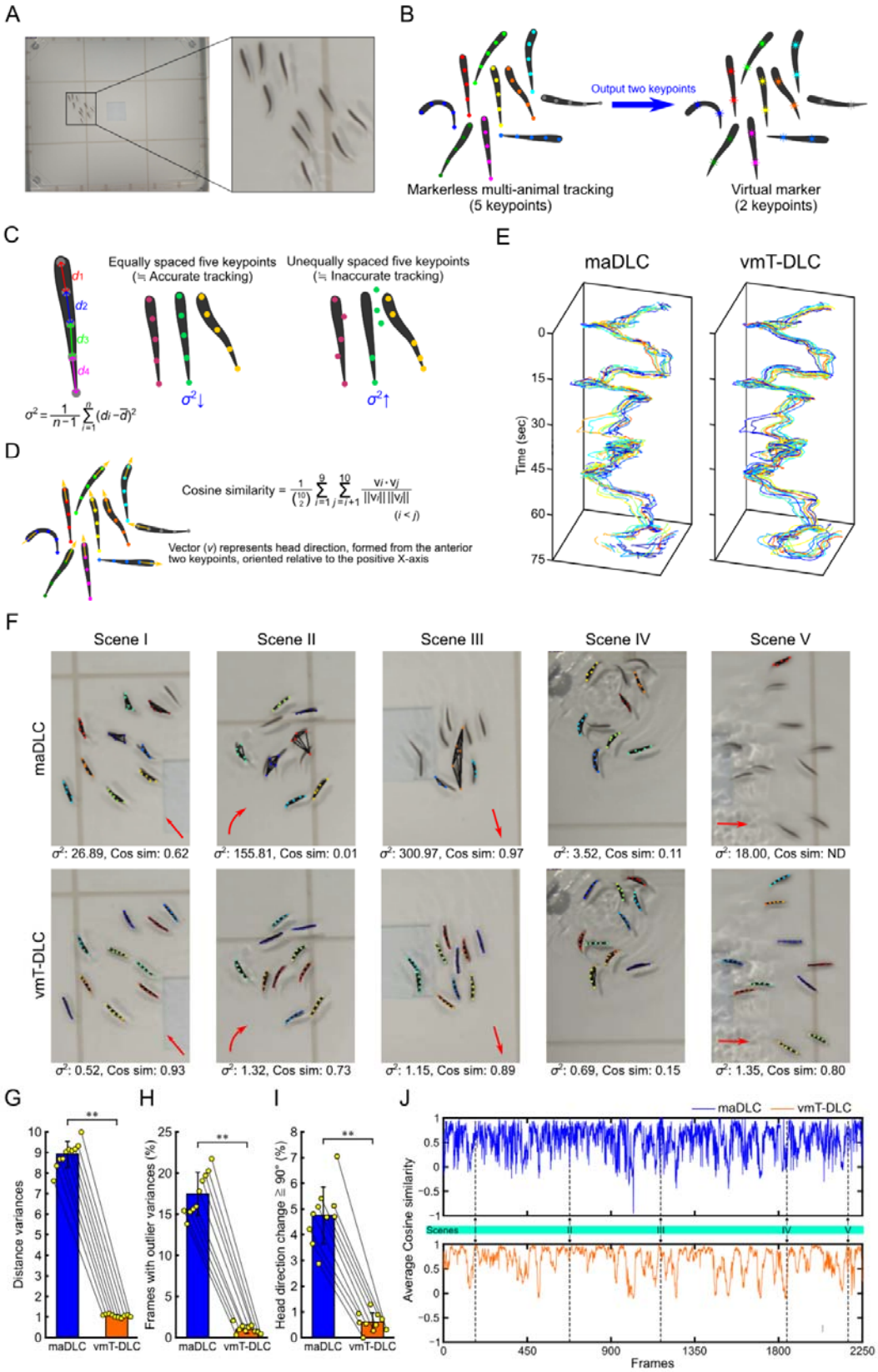
A
B
Output two keypoints
Markerless multi-animal tracking
(5 keypoints)
Virtual marker
(2 keypoints)
C
Equally spaced five keypoints
(≒ Accurate tracking)
Unequally spaced five keypoints
(≒ Inaccurate tracking)
$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$
D
Cosine similarity = $\frac{1}{\binom{10}{2}}\sum_{i=1}^{9}\sum_{j=i+1}^{10}\frac{v_i \cdot v_j}{\|v_i\|\,\|v_j\|}$ $(i < j)$
Vector (v) represents head direction, formed from the anterior two keypoints, oriented relative to the positive X-axis
E
maDLC
vmT-DLC
Time (sec)
F
Scene I
Scene II
Scene III
Scene IV
Scene V
maDLC
vmT-DLC
σ²: 26.89, Cos sim: 0.62
σ²: 155.81, Cos sim: 0.01
σ²: 300.97, Cos sim: 0.97
σ²: 3.52, Cos sim: 0.11
σ²: 18.00, Cos sim: ND
σ²: 0.52, Cos sim: 0.93
σ²: 1.32, Cos sim: 0.73
σ²: 1.15, Cos sim: 0.89
σ²: 0.69, Cos sim: 0.15
σ²: 1.35, Cos sim: 0.80
G
Distance variances
H
Frames with outlier variances (%)
I
Head direction change ≥ 90° (%)
maDLC
vmT-DLC
J
Average Cosine similarity
Scenes
Frames

Fig. 5. Comparison of markerless multi-animal pose tracking and vmTracking in multi-fish tracking.

(A) Image of the tracked school of fish. Right panel: enlarged view of the rectangle in the left panel. (B) Schematic of virtual marker creation. Five keypoints per fish were tracked, with two designated as virtual markers. (C–D) Methods for evaluating tracking. Based on tracking with equidistant keypoints, variance in distance between adjacent keypoints was used as a basic indicator of accurate tracking (larger variance indicates lower accuracy) (C). Cosine similarity (Cos sim) of head direction for ten fish was calculated using angles formed by the two foremost keypoints (D). (E) Swimming trajectories of ten fish based on the foremost (head) keypoint. (F) Example photos from five scenes and corresponding values for variance (σ^2) and Cos sim. Red arrows in the photos indicate general movement direction of the school of fish. Scene IV did not show a clear direction of movement as a group (not indicated). (G–I) Median variance for each fish (plot). (G), the number of frames showing outliers in variance for each fish based on the interquartile range method (plot); (H), the number of frames for each fish where the head direction changed by more than 90° between frames (approximately 0.03 seconds) (plot) (I), with bars and error bars representing mean ± standard deviation calculated from the medians of all ten fish. Statistical testing was performed using the Wilcoxon signed-rank test. n = 10, **: $p < 0.01$. (J) Time series changes in Cos sim. Scenes I-V correspond to Scenes I-V in (F).

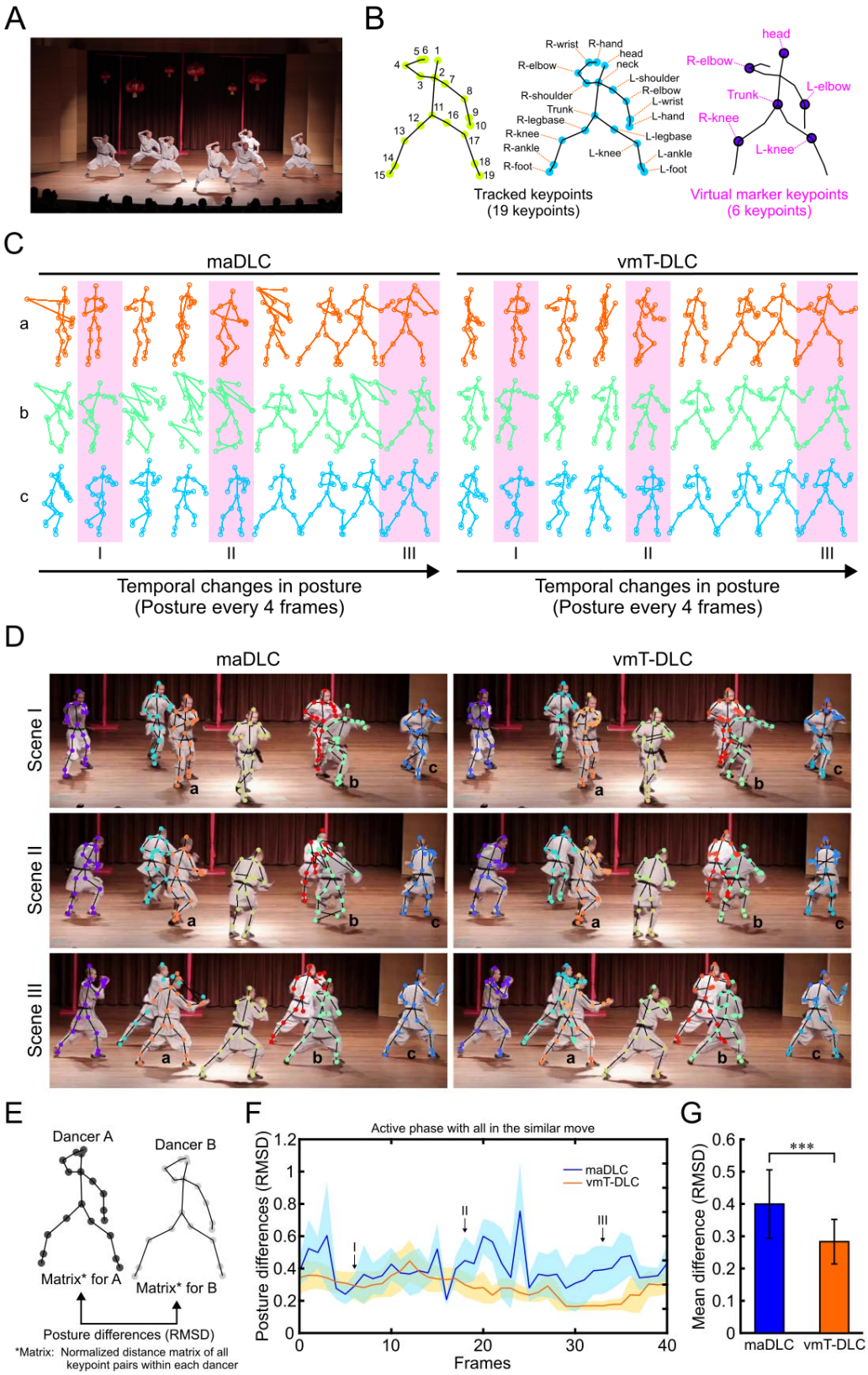
A
B
Tracked keypoints
(19 keypoints)
Virtual marker keypoints
(6 keypoints)
head
R-elbow
L-elbow
Trunk
R-knee
L-knee
C
maDLC
vmT-DLC
a
b
c
I
II
III
Temporal changes in posture
(Posture every 4 frames)
D
Scene I
Scene II
Scene III
E
Dancer A
Dancer B
Matrix* for A
Matrix* for B
Posture differences (RMSD)
*Matrix: Normalized distance matrix of all keypoint pairs within each dancer
F
Active phase with all in the similar move
maDLC
vmT-DLC
Posture differences (RMSD)
Frames
G
***
Mean difference (RMSD)
maDLC
vmT-DLC

Fig. 6. Comparison of markerless multi-animal pose tracking and vmTracking in multi-dancer tracking.

(A) Image of the video being tracked. (B) Tracked keypoints and the keypoints used as virtual markers. (C) Extraction of tracking data for three dancers from a scene where seven dancers perform similar movements over two seconds, displayed at nine points every four frames. The dancers labeled a to c correspond to dancers a to c in (D), and I to III correspond to I to III in (D) and (F). (D) Comparison of actual tracking photos for scenes I to III shown in (C). (E) Method for assessing the similarity of poses among the seven dancers. A standardized distance matrix was created for each dancer, comparing all keypoint pairs, and the root mean square deviation (RMSD) of these matrices was calculated between dancers. (F) Time series variation in the average RMSD values of all dancer pairs for each frame over two seconds, with the shaded area indicating standard deviation. (G) The overall mean and standard deviation of RMSD values calculated across all frames during the two seconds. Statistical tests were performed using paired t-tests. $n = 41$, ***: $p < 0.001$.

Supplementary figures

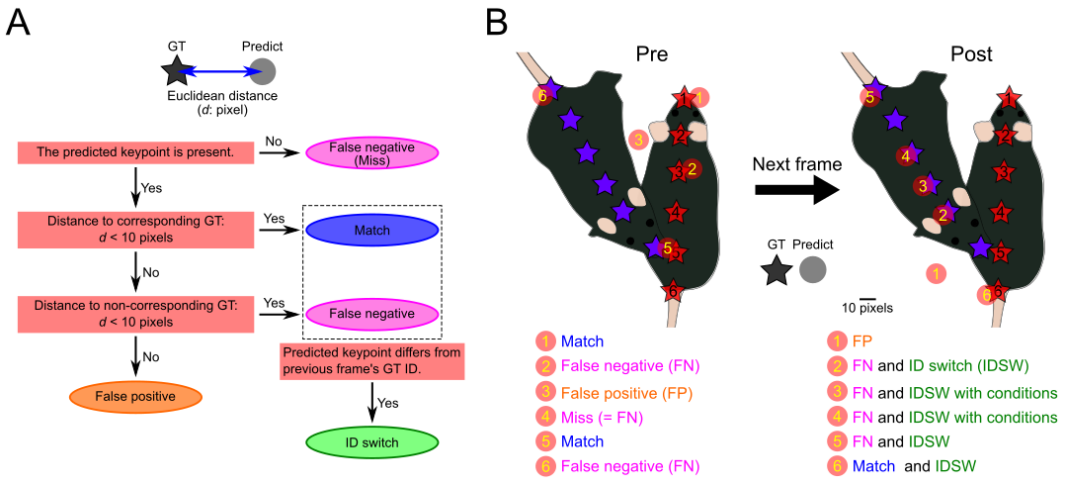

Fig. S1. Classification of predicted keypoints for tracking performance evaluation. (A) Workflow of predicted keypoint classification. If no predicted keypoint was present, it was considered a miss and treated as a false negative (FN). If a predicted keypoint was within 10 pixels of the corresponding ground truth (GT), it was considered a match. If a predicted keypoint was within 10 pixels of a non-corresponding GT, it was considered an FN because it did not predict the correct point. In cases where both the Match and FN criteria were met, the match was prioritized. If it was not within 10 pixels of any GT, it was considered a false positive (FP), predicting a non-existent point. In addition, if the ID determination of a predicted keypoint differed from the ID determination of the previous frame, then an ID switch (IDSW) occurred. (B) Schematic of classification examples. For Pre (left), 1: accurately predicts a location within 10 pixels of the corresponding GT (Match); 2: predicts a location more than 10 pixels away from the corresponding GT and within 10 pixels of a non-corresponding GT (FN), specifically a Body part mismatch FN; 3: does not predict within 10 pixels of any GT, classified as a false positive (FP); 4: shows no predicted keypoints (Miss = FN), which falls under a body part mismatch FN; 5: is within 10 pixels of both the corresponding and non-corresponding GTs, with the corresponding GT being prioritized (Match); and 6: predicts a location more than 10 pixels away from the corresponding GT and within 10 pixels of a non-corresponding GT (FN), classified specifically as a body part match/ID mismatch FN. This involves predicting a location with a different ID. However, since ID switches (IDSW) are based on changes from the previous frame, it does not qualify as an IDSW here. For Post (right), 1: fails to predict within 10 pixels of

any GT (FP); 2: predicts a location more than 10 pixels from the corresponding GT and within 10 pixels of a non-corresponding GT (FN), specifically a body part match/ID mismatch FN, and it also predicts a GT with a different ID from the previous frame (IDSW); 3 and 4: predict locations more than 10 pixels from the corresponding GT and within 10 pixels of a non-corresponding GT (FN), classified specifically as body part match/ID mismatch FN. They are predicting GTs with different IDs. However, in the previous frame, 3 is classified as a false positive (FP) and 4 as an FN due to a Miss. If the predicted IDs had changed from even earlier frames, this situation would be classified as an ID switch (IDSW); 5: predicts a location more than 10 pixels away from the corresponding GT and within 10 pixels of a non-corresponding GT (FN), specifically a Body part mismatch FN, and it also predicts a GT with a different ID from the previous frame (IDSW); and 6: accurately predicts within 10 pixels of the corresponding GT (Match), and it also predicts a GT with a different ID from the previous frame (IDSW).

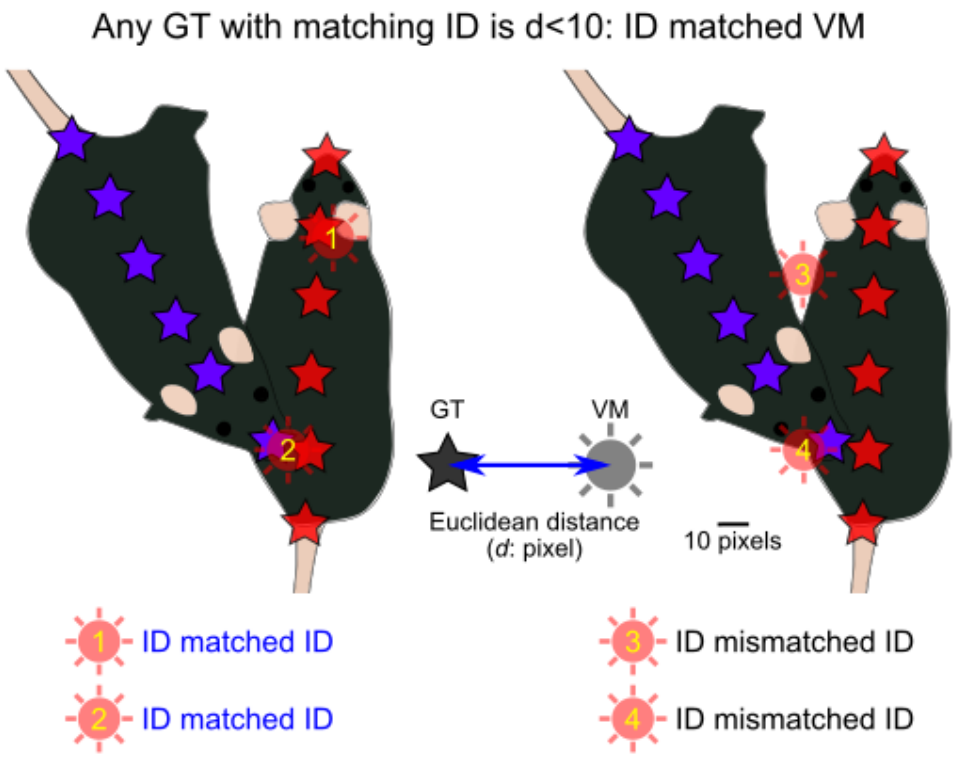

Fig. S2. Accuracy of virtual markers based on ID matching.
The accuracy of the virtual markers was based on the distance (*d*: pixels) between the virtual marker points indicated by radial circles and the ground truth indicated by stars. If any ground truth keypoints with the same ID as the virtual marker were within 10 pixels of the corresponding virtual marker, the virtual marker was considered to have an ID match (illustrated in the Fig. as A and B). Conversely, if no ground truth keypoints with the same ID as the virtual marker were within 10 pixels, the virtual marker was considered to have an ID mismatch (illustrated in the Fig. as C and D). Additionally, if virtual markers were absent due to missing predictions in markerless multi-animal pose tracking, these were treated as instances of ID mismatch, as they represent missing virtual markers that should have been present. The proportion of ID-match virtual marker points for each scene was calculated and used for the correlation analysis with the accuracy of vmTracking.

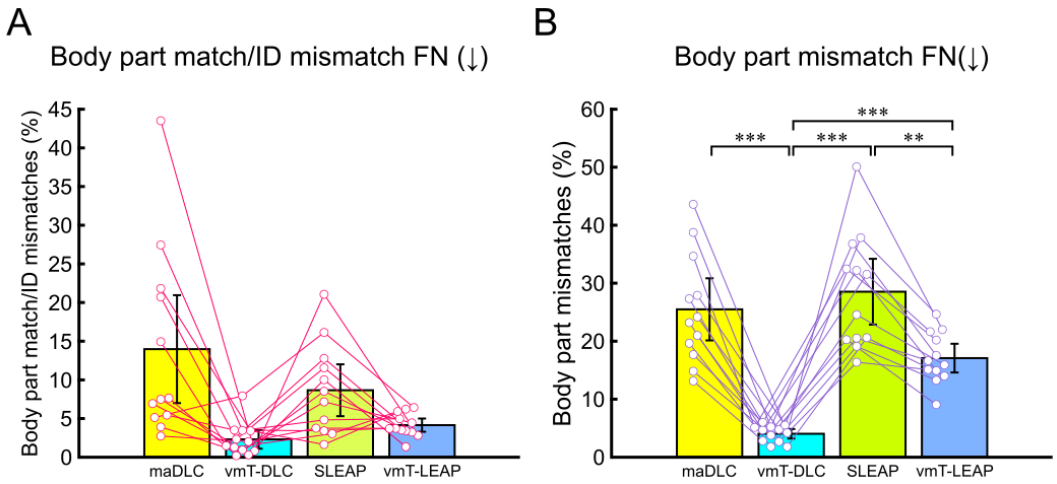

Fig. S3. Evaluation through detailed classification of false negatives in five mice tracking.

(A–B) For the 12 occlusion-crowded scenes sampled (1,800 frames in total), a comparison was made with the manually created ground truths. This comparison examined two types of false negatives: (A) cases where the predicted ID does not match, but the body parts do match (Body part match/ID mismatch FN), and (B) cases where the predicted body parts do not match, including missing parts (Body part mismatch FN). These were compared for multi-animal DeepLabCut (maDLC), vmTracking with single-animal DeepLabCut (vmT-DLC), Social LEAP (SLEAP), and vmTracking with LEAP (vmT-LEAP). Arrows indicate better values. Data represent mean ± 95% confidence interval (n = 12 scenes). Statistical tests were performed using repeated-measures ANOVA and Bonferroni correction. **: $p < 0.01$, ***: $p < 0.001$.

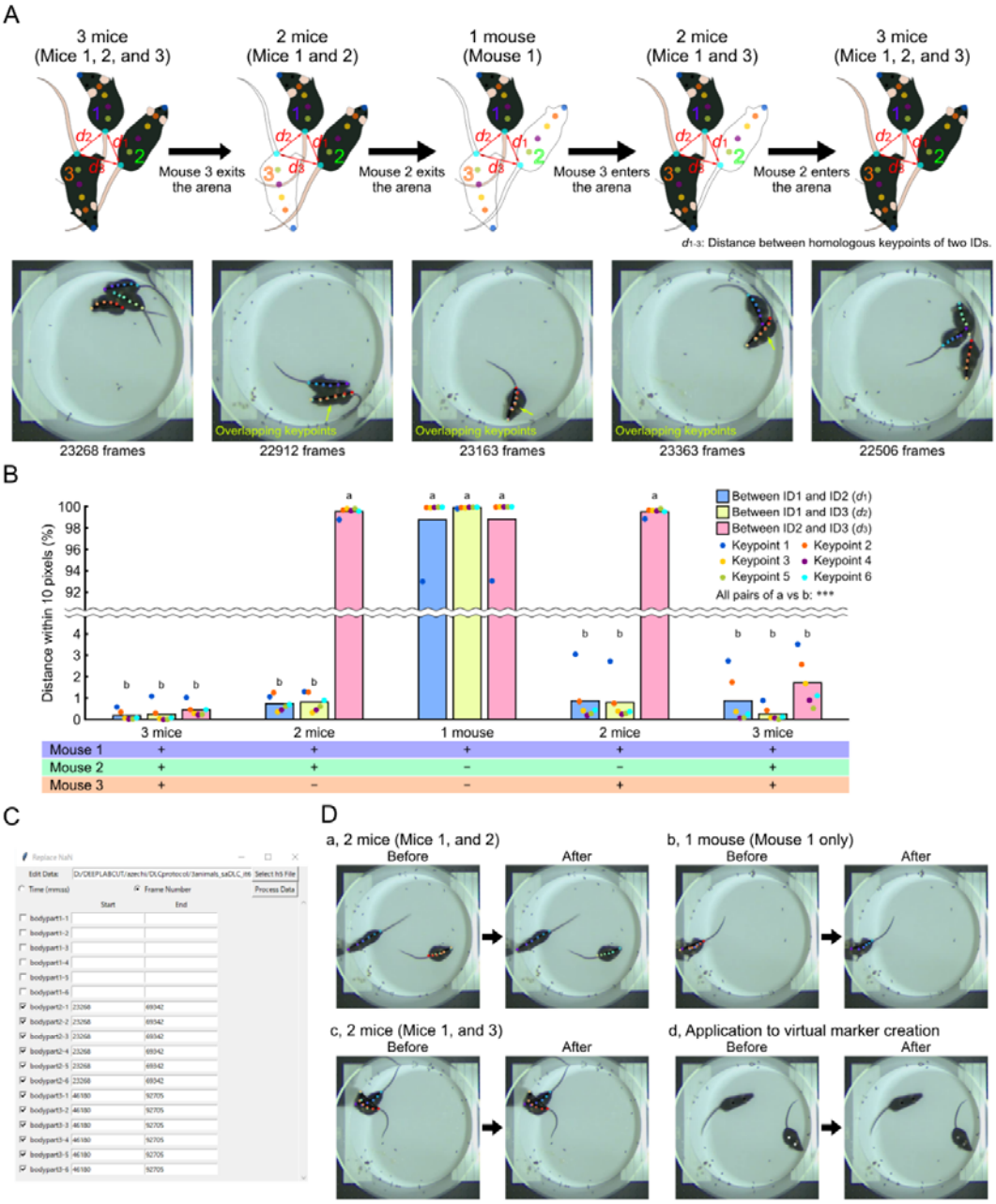

Fig. S4. Tracking when the number of individuals in the video frame changes. (A) Schematic diagram (top) of an experiment that altered the number of mice in the video frame by introducing or removing mice from the arena, and examples of tracking photos during this procedure (bottom). To verify tracking of other mice with leftover keypoints due to the absence of some mice, distances (*d*: pixels) between identical keypoints for three IDs were calculated for each keypoint. (B) If the distance between identical keypoints was within 10 pixels, the two ID keypoints were considered

overlapping, and the frequency of such frames was calculated for each keypoint and ID pair combination. The diagram illustrates changes across different experimental conditions, with plots indicating the frequency for each keypoint and bars representing each ID pair. The '+' below the diagram indicates the presence of mice in the arena, while '−' indicates their absence. Statistical tests were conducted using mixed two-way ANOVA with ID pair as the between-subject factor and experimental conditions (manipulations of mouse numbers) as the within-subject factor, followed by post hoc tests with Bonferroni correction. $n = 6$, ***: $p < 0.001$. (C) Photo of the GUI used to remove specified data from the coordinate data file (h5 format) obtained from DLC, making it NaN. The process is executed by specifying the keypoint and range to be removed (combination of fps and time, or frame number). (D) Examples of tracking video images before and after data processing through the GUI described in (C). (Da–c) Markers for Mouse 1 (blue), Mouse 2 (green), and Mouse 3 (red) overlap variously depending on the condition (Before), but after removing the data of the absent mice, a video is created that shows no overlap and refined tracking by correct IDs (After). (Dd) Even in markerless multi-animal pose tracking, markers occasionally overlap. Removing such overlaps allows for the creation of correct virtual marker videos. In this example, the white data obscuring the correct gray virtual marker is removed.

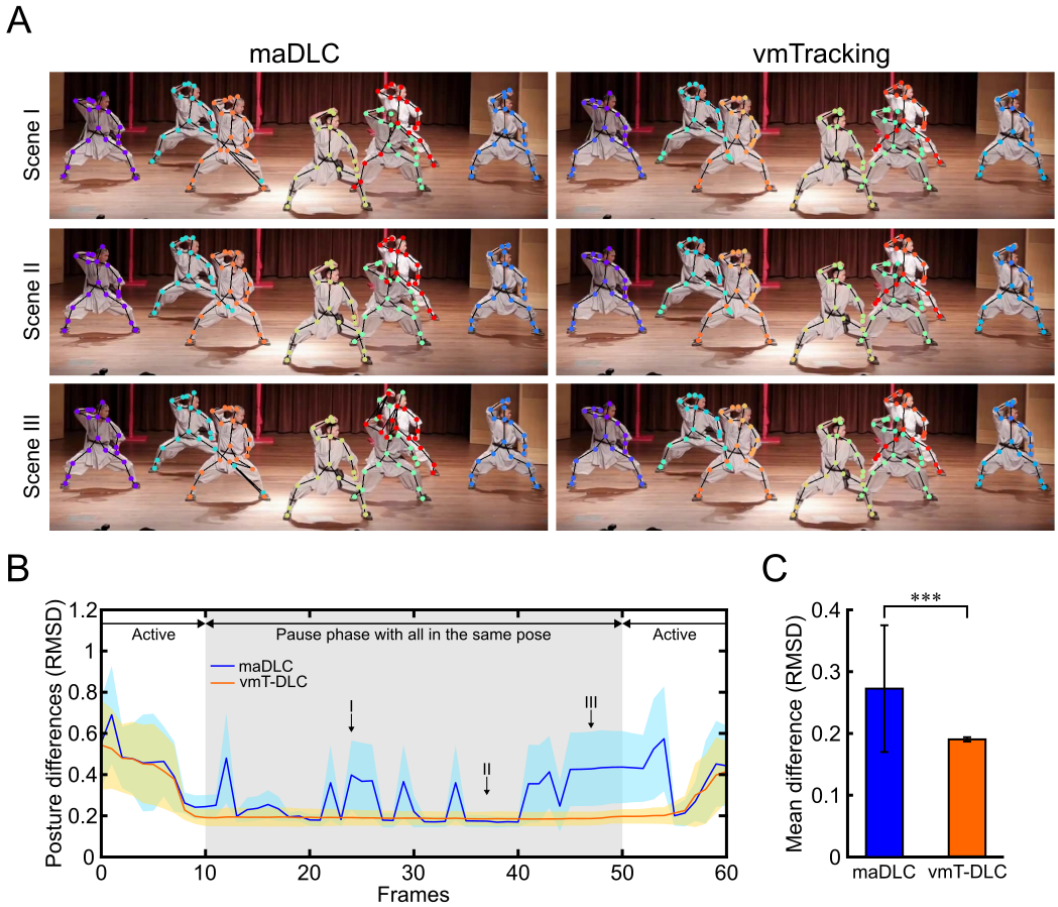

Fig. S5. Tracking of multiple dancers in a pause state.

(A) Three examples of tracking scenes where seven dancers are posed similarly and remain stationary for two seconds. (B) Time series variation of RMSD over three seconds, including approximately two seconds of pause scenes. The shaded area along the plot line indicates standard deviation. The gray shaded area in the Fig. represents the pause scenes. I to III correspond to scenes I to III in (A). (C) Comparison of RMSD between maDLC and vmT-DLC during the two-second stationary scenes. Data are presented as mean ± standard deviation. Statistical tests were conducted using paired t-tests. n = 41, ***: $p < 0.001$.

Supplementary Videos

Supplementary Videos S1

Video clips demonstrated vmTracking applied across various species and conditions.

Supplementary Videos S2

The comparative video clip demonstrated the performance of maDLC (A) and vmT-DLC (B) in tracking five mice. The video of vmT-DLC had virtual markers excluded. The video showed tracking results over five seconds at 0.1x speed. In the maDLC footage, there were identity switches indicated by label colors and missing predicted labels. In contrast, vmT-DLC consistently maintained accurate tracking throughout.

Supplementary Videos S3

The comparative video clip demonstrated the performance of maDLC (A) and vmT-DLC (B) during the tracking of three mice in an occlusion-crowded scene. The video of vmT-DLC had virtual markers excluded. Beyond the five-second mark in the video, the mouse labeled in red was partially obscured by the mouse labeled in green. In this scene, while the prediction of the mouse labeled in red was almost missing in maDLC, vmT-DLC maintained reasonable tracking.

Supplementary Videos S4

The comparative video clip demonstrated the performance of maDLC (A) and vmT-DLC (B) during the tracking of three mice in a non-occlusion-crowded scene. The video of vmT-DLC had virtual markers excluded. In scenes without occlusion or crowding, both methods maintained consistent and accurate tracking.

Supplementary Videos 5

The comparative video clip demonstrated the performance of maDLC (A) and vmT-DLC (B) in tracking a school of ten fish. The video of vmT-DLC had virtual markers excluded. The video showed tracking results over 15 seconds at 0.5x speed. In maDLC, there were occurrences of missing predicted labels and some keypoints predicting different fish, whereas vmT-DLC generally maintained stable tracking of each fish.

Supplementary Videos 6

The comparative video clip demonstrated the performance of maDLC (A) and vmT-DLC (B) in tracking seven dancers. The video of vmT-DLC had virtual markers excluded. In both maDLC and vmT-DLC, some keypoints occasionally predicted different dancers. However, while maDLC showed instances

where entire sets of keypoints switched to tracking another dancer, vmT-DLC did not exhibit such switches.